The Structure and Evolution of the Protein Genotype-Phenotype Map

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Evandro Ferrada
aus
Chile

Promotionskomitee:
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Frédéric Allain
Prof. Dr. Homayoun Bagheri
Prof. Dr. Amedeo Caflisch

Zürich, 2011

To GP

La cruel blancura de la eternidad
hace que la luz huya de sí misma…
…algo nos recuerda la verdad
que amamos antes de conocer.
Jorge Teillier

Summary

By exploring the genotype–phenotype map of macromolecules researchers aim to propose possible explanations for questions such as how RNA and proteins have emerged and changed through evolution, and what properties help to explain the diversity of molecular functions and their adaptation to different environments.

Despite its complexity, the sequence–structure relationship of macromolecules represents the simplest example of a genotype–phenotype map. In this map, genotypes correspond to sequences and phenotypes to structures or functions.

Two main areas of research have facilitated the study of genotype–phenotype maps of macromolecules. First, the accumulation of sequence and structure data from genomics, metagenomics, and structural genomics initiatives. Second, the development of computational models that allow the prediction of three-dimensional structures and the exhaustive exploration of simple models of biopolymers.

This dissertation is mainly concerned with the protein genotype-phenotype map. I begin with a revision of the literature and then, explore a few questions related to the organization and evolution of the protein genotype–phenotype map (chapter 1). The second chapter in this dissertation is entitled "a comparison of genotype-phenotype maps for RNA and proteins" (Ferrada and Wagner, submitted). In this chapter, I compare properties of simple models of proteins and RNA. I find that many fewer protein molecules than RNA molecules fold, but they fold into many more structures than RNA. In consequence, protein phenotypes have smaller genotype networks whose member genotypes tend to be more similar than those of RNA phenotypes. Neighborhoods in sequence space of a given radius around an RNA molecule contain more novel structures than for protein molecules. I compare this property with evidence from natural RNA and protein molecules and conclude that RNA genotype space may be more conducive to the evolution of new structure phenotypes.

Chapter 3 in this dissertation is entitled “Protein robustness promotes evolutionary innovations on large evolutionary time scales” (Ferrada and Wagner 2008). In this work I explore the interplay between structural robustness and functional innovations in proteins. I study protein domains conserved in all extant organisms and different functional classifications of proteins and show that more robust proteins have a greater capacity to produce functional innovations.

In chapter 4, entitled “Evolutionary innovations and the organization of protein functions in genotype space” (Ferrada and Wagner 2010), I show that different neighborhoods of genotype space contain proteins with very different functions. This property both facilitates evolutionary innovation through exploration of a genotype network, and constrains the evolution of novel phenotypes. I show that the space of protein functions is not homogeneous, and different genotype neighborhoods tend to contain a different spectrum of functions, whose diversity increases with increasing distance of these neighborhoods in sequence space.

In the last chapter of this dissertation, I introduce mathematical formalisms related to the concept of genotype space and explore some of its properties. I show that consensus sequences of protein neutral networks distribute randomly in sequence space at distances that scale according to amino acid alphabet size. Additionally, I show how these observations can be used to deepen our understanding of the evolution of the protein genotype-phenotype map. I finally propose a simple model that aims to reconcile the extensive size variation observed in natural proteins with the genotype space concept.

Zusammenfassung

Mit der Erforschung der Genotyp-Phänotyp-Beziehung von Makromolekülen versuchen Forscher Einblick zu gewinnen in die Entstehung und Evolution von RNA und Proteinen. Darüber hinaus wird versucht, diejenigen Eigenschaften dieser Moleküle zu ergründen, welche zu ihrer funktionellen Diversität und Anpassung an unterschiedliche Umgebungen beitragen.

Trotz seiner Komplexität ist die Sequenz-Struktur-Beziehung von Makromolekülen das einfachste Beispiel einer Genotyp-Phänotyp-Verknüpfung. Dies beruht auf einer Vielzahl von empirischen und theoretischen Informationen, die in diesem Feld zusammengetragen wurden. In diesen Verknüpfungen entspricht der Genotyp der Sequenz und der Phänotyp der Struktur oder Funktion.

Die folgenden zwei Wissenschaftsfelder haben die Forschung an der Genotyp-Phänotyp-Beziehung von Makromolekülen vereinfacht: Erstens kam es in der jüngsten vergangenheit zu einer Akkumulation von Sequenz- und Strukturdaten durch Genomik, Metagenomik, und strukturbasierter Genomik. Zweitens erlaubte die Entwicklung von Modellberechnungen sowohl die Vorhersage dreidimensionaler Strukturen als auch die tiefgehende Untersuchung an einfacher Biopolymermodelle.

Diese Dissertation beschäftigt sich primär mit der Genotyp-Phänotyp-Beziehung von Proteinen. Ich werde mit einem kurzen Literaturrückblick beginnen und diskutiere danach einige Fragen zur Organisation und Evolution der Genotyp-Phänotyp-Beziehung von Proteinen. Die erste Studie meiner Dissertation trägt den Namen „Ein Genotyp-Phänotyp Beziehungsvergleich für RNA und Proteine" (Ferrada und Wagner, eingereicht). In diesem Abschnitt vergleiche ich die Eigenschaften von einfachen Protein- und RNA-Modellen. Dabei fand ich heraus, dass sich viel weniger Protein- als RNA-Moleküle falten. Jedoch stellte ich ebenfalls fest, dass trotz der geringeren Anzahl sich faltendender Proteine, diese eine höhere Strukturdiversität besitzen. Daraus ergibt sich, dass Protein-Phänotypen kleinere und ähnlichere Genotyp-Netzwerke haben als RNA-Phänotypen. RNA-Moleküle innerhalb eines definierten Radius in einen RNA Genotype zeigen dagegen neuartigere

Strukturen als Proteine. Ich vergliche diese Eigenschaften mit Beobachtungen an natürlichen RNAs und Proteinen und schlosse daraus, dass RNA-Genotypen zugänglicher für evolutionäre Veränderungen des Struktur-Phänotyps sind.

Die zweite Studie dieser Dissertation mit dem Titel "Protein Stabilität fördert langfristig die evolutionäre Innovation" (Ferrada und Wagner 2008) beschäftigt sich mit dem Zusammenspiel von struktureller Stabilität und funktioneller Erneuerung in Proteinen. Dabei untersuchte ich in allen Organismen konservierte Proteindomänen und deinen unterschiedliche Funktionen um zu zeigen, dass je robuster ein Protein ist, desto grösser ist sein Vermögen funktionelle Erneuerungen zu erzeugen.

In einer dritten Studie mit dem Namen "Evolutionäre Innovationen und die Organisation von Proteinfunktionen im Genotyp-Raum" (Ferrada und Wagner 2010), zeigte ich, dass Proteine mit unterschiedlichen Genotypnachbarschaften sehr unterschiedliche Funktionen beinhalten. Diese Eigenheiten bewirken, dass sowohl evolutionäre Innovationen durch das Ausnutzen ganzer Genotyp-Netzwerke erleichtert werden, als auch die Evolution neuartiger Phänotypen gezügelt wird. Ich konnte weiterhin zeigen, dass der Raum der Proteinfunktionen nicht homogen ist, und dass unterschiedliche Genotyp-Nachbarschaften dazu tendieren ein unterschiedliches Funktionsspektrum zu beinhalten, dessen Diversität mit zunehmendem Abstand zu benachbarten Sequenzen ansteigt.

Im letzten Kapitel dieser Dissertation erarbeite und untersuche ich mathematische Formalismen, die mit dem Konzept des Genotyp-Raums in Verbindung stehen. Ich zeige, dass neutrale Netzwerke sich zufällig im Sequenzraum verteilen, und dass dabei der Abstand von der Grösse des Aminosäurenalphabets abhängig ist. Darüber hinaus zeige ich wie diese Beobachtungen genutzt werden könnten, um unser Verständnis von Protein-Genotyp-Phänotyp-Beziehungen zu vertiefen. Schliesslich schlage ich ein einfaches Modell vor, das das Ziel verfolgt die in natürlichen Proteinen beobachteten umfangreichen Grössenvariationen mit dem Genotyp-Raum-Konzept zu harmonisieren.

Contents

1. Introduction

1.1 Historical background

The modern molecular biology era began with the discovery of DNA as the underlying substrate of genetic information. As a result, proteins, which had previously been thought to accomplish this function, were confined to the role of the 'working machinery' of the cell. The central dogma of molecular biology expressed this paradigm-shift explicitly, representing proteins as the final product of the flow of genetic information (Crick 1970). The beginning of the molecular biology era regarded proteins as static objects whose structures and functions needed description and classification. This was the attitude that during the 1950s inspired the experiments of Christian Anfinsen on ribonuclease (Sela et al 1957; Anfinsen 1973) and the solution of the first protein crystal structure by John Kendrew (Kendrew et al 1958), giving birth to structural biology as a discipline.

Anfinsen's experiment revealed that proteins were chemical objects governed by physical laws. Although this conclusion may seem obvious to us today, it had two important consequences: first, it made the knowledge of the chemistry of simple molecules applicable to proteins; and second it helped to reinterpret protein folding as a thermodynamic process. According to Anfinsen's *thermodynamic hypothesis*, protein sequences contain all the necessary information to give rise to their native structures, and the structure they adopt corresponds to a global minimum of free energy (Anfinsen 1973). Additionally, during the same decade, the solution of the first protein crystal structure revealed that proteins usually encompass thousands of atomic interactions. It also showed that the atomic organization of proteins presents regularities that were later referred to as secondary and tertiary structures (Kendrew et al 1958).

During the next years, research interests focused on *protein folding*, the process through which a particular sequence adopts its native structure. The immense number of degrees of freedom that a polypeptide chain possesses during folding led Cyrus Levinthal (Levinthal 1968) to propose that a polypeptide would need a very large amount of time to explore the possible space of conformations. However, it was also known that proteins fold inside the cell in fractions of a second. These contradictory observations later became

known as *Levinthal's paradox*. Levinthal proposed that proteins follow a series of conformational changes, which he envisioned as tunnels through the space of conformations, or *folding pathways* (Levinthal 1968) (Figure 1.1).

Simple models of proteins, originally proposed to study polymer dynamics, were used to explore folding (Lau and Dill 1989; see Section 1.6). These models are based on short polymers composed of only two types of monomers (hydrophobic, H; and polar, P). Polymers fold on a grid using a set of discrete spatial movements, whose simplicity compared to natural proteins allows a detailed exploration of the folding dynamics. Simple models revealed folding as a complicated process of energetic interaction among amino acids (Dill et al 1995). It was only during the 1990s that a '*new view*' of protein folding emerged based on the concept of an *energy landscape* (Frauenfelder et al 1991, Bryngelson et al 1995). Dill and Chan summarized theoretical and empirical findings by proposing that instead of following individual pathways, folding corresponds to a parallel process, where a protein diffuses through conformational space to reach its native structure (Dill et al 1995; Dill and Chan 1997) (see Figure 1.1).

During the decade between 1960 and 1970, classical population genetic principles were applied to collections of DNA and protein sequences, giving rise to the field of *molecular evolution*. The comparison of multiple DNA coding sequences from different species led Zuckerkandl and Pauling to propose the *molecular clock hypothesis* according to which amino acids substitutions

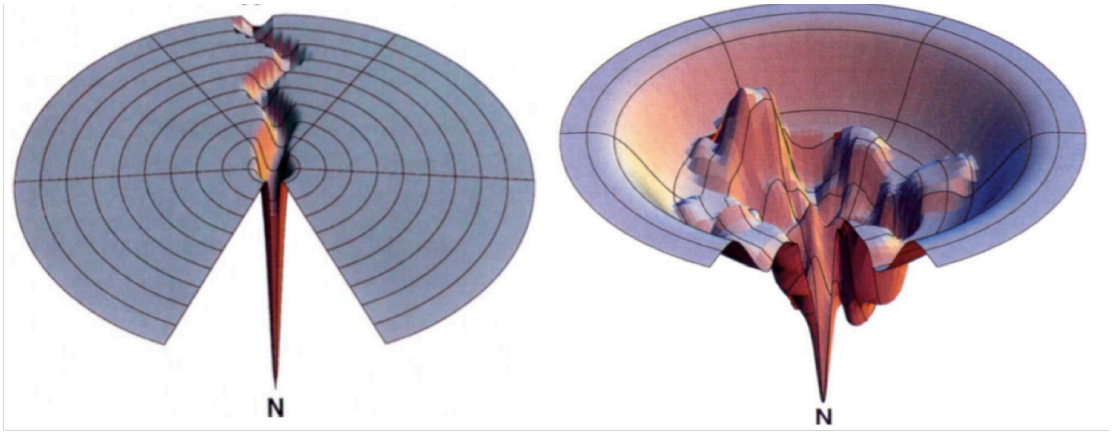


Figure 1.1. *Folding pathways, energy landscapes and the new view of protein folding*. Left, a folding pathway as envision by Levinthal (Levinthal 1968). Right, the new view of protein folding. Figure modified from (Dill and Chan 1997). Reprint with permission from Macmillan Publishers Ltd: Nat Str Biol, copyright (1997).

accumulate in coding sequences at a constant rate. Therefore sequence analysis can help date the fossil record (Zuckerkandl and Pauling 1965). Influenced by this idea and by his own theoretical work Kimura proposed the *neutral theory of molecular evolution,* after noticing that most of the point mutations experienced by a protein molecule may have no impact on its fitness (Kimura 1968; 1983), meaning that, on average, a large fraction of mutations leave protein structures unaltered. This idea generated an important debate that still hovers in the current molecular evolution literature, namely the relative importance of neutral evolution (*neutralism*) versus the role of natural selection (*selectionism*) in molecular evolution (Nei 2005).

In 1965 Margaret Dayhoff pioneered the field of *bioinformatics* by compiling an atlas of known protein sequences. The comparison of sequences from different species led her to the estimation of transition probabilities between amino acids, a key tool for the analysis of protein evolution that became later known as an amino acid *substitution matrix* (Dayhoff 1978). In addition, she coined the concept of a *protein family* as a set of proteins that share a common evolutionary ancestor (Dayhoff 1976).

Since 1980, the accumulation of protein data has made comparative studies between sequences and structures possible. One of the first observations was that of the predominant conservation of structures compared to sequences during evolution (Chothia and Lesk, 1986; Ptitsyn and Volkenstein 1986). Sequence–structure comparisons showed that even with a low sequence similarity of around 20 to 30 percent two proteins typically share the same structure (Chothia and Lest 1986). This observation was the basis for the comparative approach to *structural modeling*, which is the currently most successful technique to study the space of protein structures (Sali and Blundell 1993; Marti-Renom et al 2000).

The creation of the first repositories devoted to storing and classifying protein structural motifs or domains revealed important information about proteins (Orengo et al 1994, Murzin et al 1995). A structural domain (or fold) is usually defined as a compact, autonomous folding polypeptide sequence that also forms a unit of evolution (Branden and Tooze 1999; Murzin et al 1995). The analysis of hundreds of proteins into these structural units demonstrated that

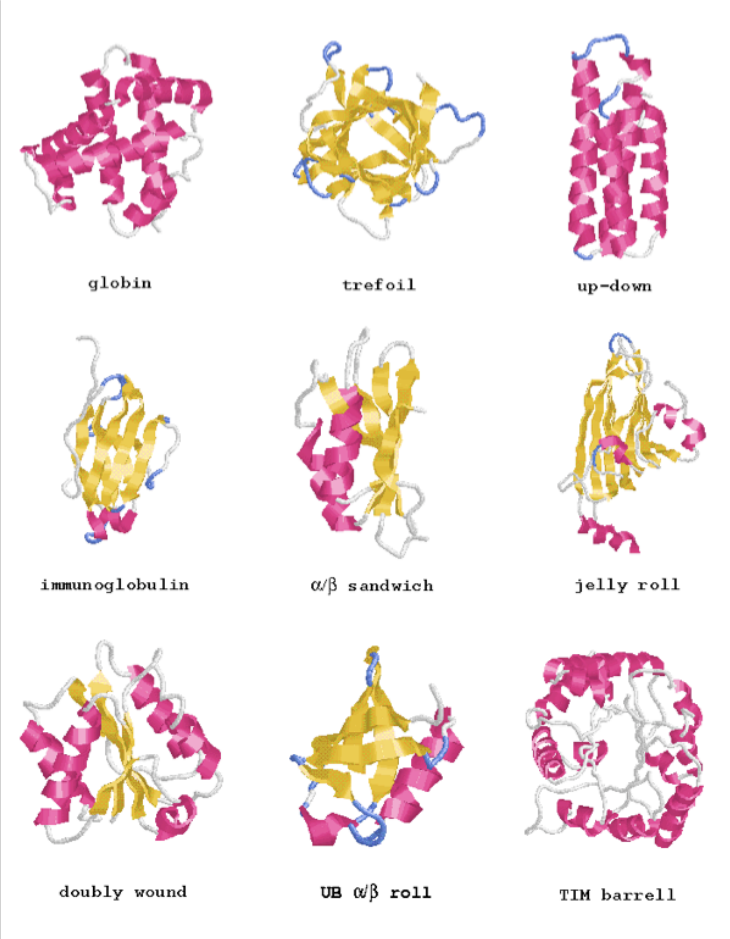


Figure 1.2. *Common protein folds*. Nine common folds identified by the first large scale analysis of protein structures (Orengo et al 1994). Folds are: globin, trefoil, up-down, immunoglobulin, alpha/beta sandwich, jelly roll, double wound, UB alpha/beta roll and TIM barrel. Reprinted with permission from Macmillan Publishers Ltd: Nature, copyright (1994).

while some folds are highly conserved throughout the protein world, the vast majority are rare (Orengo et al 1994) (Figure 1.2).

Through the accumulation of protein structures and their classification into a limited set of conserved folds, an increasing amount of research could be devoted to study the origin and evolution of proteins (Grishin 2001; Lupas et al 2001). Recent developments in sequencing technology, have further facilitated the study of macromolecular evolution. Relatedly, large-scale *in vitro* experiments have provided information on the interaction of protein properties during evolution, such as stability and functionality (Bloom et al 2006; Aharoni et al 2005). These experiments have also asked how proteins adapt to new environments (Bloom et al 2007; Bloom and Arnold 2009).

The concept of the *protein universe* refers to all proteins that exist or have ever been explored by nature (Ladunga 1992; Holm and Sander 1996; Levitt 2009). As of June 2011, there are ~15 million sequences (UniProtKB/SwissProt; The UniProt Consortium 2010) and ~70,000 protein crystal structures deposited in the Protein Data Bank (PDB), the largest internet repository of structural biology data (Berman et al 2000). Although these quantities represent only a tiny fraction of all possible protein sequences, complementary analyses suggest that our knowledge of the structural universe of proteins might be nearly complete (Zhang et al 2006; Levitt 2009).

Structural biology, at the age of fifty, faces challenges that go beyond the structural and functional characterization of new proteins. Our current knowledge of proteins from sequences to functions cries out for a common unified framework that can help understand the origins and evolution of proteins.

Such a framework should help explore the limits of protein space in terms of sequences and structural forms; the origin and evolution of new forms and functions, be it *de novo* or from previously existing proteins; the interplay between sequence and structure during protein adaptation; the role of intrinsic physicochemical properties that hinder or foster protein evolution at the sequence and structure level; and in the era of synthetic biology, the *de novo* design of proteomes.

Furthermore, the sequence-structure relationship in proteins becomes of general biological relevance when it is seen as a special case of a generic *genotype–phenotype map* (Alberch 1991). Interestingly, the genotype–phenotype map echoes the *protein folding problem* announced at the beginning of the molecular biology era (Anfinsen et al 1973). After fifty years of research, proteins represent the empirically best explored and simplest natural example of a genotype–phenotype map.

Some of the most important questions in evolutionary biology regard the dynamic interplay between the information encoded in the genotype and how this information produces a phenotype. In simple terms, the concept of a genotype-phenotype map encapsulates how variation at the level of the genotype affects the phenotype (Wagner and Altenberg 1996). The importance

of the genotype–phenotype map is that it encapsulates the raw material of natural selection, and therefore understanding its organization will shed light on the potentialities and intrinsic limits of evolution.

1.2 Sequences and the amino acid alphabet.

Proteins are polymers composed of amino acids. The number of different monomers that composes a polymer is called the *monomer alphabet size*. In the case of proteins, the amino acid alphabet size ($|\boldsymbol{A}|$) is twenty. Amino acids are joined in a head to tail manner by peptide bonds to form strings called amino acid sequences. The length of a protein (L) is the number of amino acids that composes it. Protein length varies extensively. For instance the shortest and longest known protein have 2 (uniprot accession number: P83570) and ~35,000 amino acids (uniprot accession number: A2ASS6), respectively (The UniProt Consortium 2010). The average length of a protein is 130 amino acids and 60 percent of known proteins have between 100 and 400 amino acids. The collection of all possible sequences with a given $|\boldsymbol{A}|$ and L is known as *sequence space* (see section 5.1).

As of June 2011, the UniProt/TrEMBL sequence database (The UniProt Consortium 2010) of proteins contained 15.4 million sequences, including data from over 2,400 completely sequenced genomes (Sayers et al 2011). An important contribution to our knowledge of sequences comes from metagenomics projects. The aim of these projects is to sample sequences from unexplored species that are difficult to cultivate under laboratory conditions (Handelsman 2004). During the last five years the number of sequences has increased sixfold and is currently increasing at a rate of 1 million sequences per 3.2 months (The UniProt Consortium 2010), partly due to the success of these projects.

There is a non-uniform distribution of amino acid frequencies observed in natural proteins, that is, some sequences are enriched in some amino acids and depleted of others. For instance the most and least common amino acids are *leucine* and *tryptophan*, which account for as much as 9.6 and as little as 1.0 percent of amino acids in known proteins, respectively. In addition to the non-uniform distribution of amino acids, sets of functionally or evolutionary related

proteins usually possess characteristic amino acid compositions that deviate from the non-uniform amino acid distribution of all known proteins.

1.3 Protein families

Since the seminal work of Margaret Dayhoff, we know that at the sequence level, proteins can be grouped into families (Dayhoff 1976). A protein family is a set of sequences with a high degree of similarity, and because of that a common evolutionary origin is usually assumed.

Based on methods of sequence comparison, it is possible to distinguish between two general types of protein family classifications. The first are classifications based solely on sequence relatedness. Among the simplest methods to study protein relatednesss are *pairwise sequence alignment algorithms* (Waterman 1995). An alignment compares the consecutive positions of amino acids sites along two sequences. Equivalent positions along the alignment are occupied by amino acids that are structural and/or functionally related. Efficient, well-known algorithms are based on dynamic programming (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Altschul et al 1990). An alignment can also be obtained for more than two sequences. When this is the case the comparison is called a *multiple sequence alignment* (MSA). There are two main indices of relatedness derived from a sequence alignment. First, *sequence identity* refers to the fraction of equivalent sites that are occupied by the same amino acid along the alignment. Second, *sequence similarity* takes into account amino acids that although not identical, possess certain degree of physical and chemical relatedness, such as size and hydrophobicity. The similarity between amino acids is usually encapsulated in *amino acid substitution matrices* that are derived from large-scale comparisons of evolutionary related sequences (Dayhoff 1978). Relatedness has an intrinsic *detection limit* or *random threshold* that is determined by amino acid alphabet size. The random threshold κ of sequence relatedness is defined as:

$$\kappa = \frac{1}{|\boldsymbol{A}|} \tag{1.1}$$

Where $|\boldsymbol{A}|$, correspond to the amino acid alphabet size. Sequence relatedness can also be expressed in terms of sequence distance or dissimilarity (1-κ). The

observed κ in natural proteins corresponds to approximately 25 percent of sequence identity (Rost 1999). In general, when two sequences have amino acid identity higher than 25 percent, they are called *homologous*. Homologous sequences detected in different species are called *orthologous* and those detected in the same organism *paralogous* (Li 1997).

Although there is no general consensus on the sequence relatedness used to classify two proteins as part of the same family, commonly used thresholds of sequence similarity range between 40 to 60 percent (Kriventseva et al 2001; Kunin et al 2005). The challenge of grouping proteins according to a sequence similarity threshold relies on the clustering of a huge amount of data. The advantage is simplicity and the use of unsupervised classification procedures (Silverstein et al 2001).

A second type of protein family classification method is based on *sequence profiles* (Sonnhammer et al 1998). Sequence profiles aim to detect sequence patterns or sets of similar amino acids distributed along a multiple sequence alignment. A sequence profile is described by the probability, p_i^k, of finding amino acid *i* at position *k* of a multiple sequence alignment. The profile is usually represented as a matrix of dimensions *i* x *k*, where *k* is the sequence alignment length and each column corresponds to p_i, the frequency distribution of amino acids in the alignment. The first sequence profile method constructed was PSI-BLAST (Altschul et al 1997).

An important advantage of sequence profiles is their independence from a similarity threshold definition. The distant homology detection provided by sequence profiles usually captures sequence signals that are essential to conserve a protein fold (ie. *structural determinants*).

Due to their simple definition, protein families represent the easiest way of organizing protein sequences. That is why much of the interest in describing the general organization and evolution of protein space has focused on the characterization of protein families (Todd et al 2001; Kunin et al 2005).

Currently there are around 15,000 protein families. Among them around 50 percent have at least one known representative structure (Sonnhammer et al 1998; Geer et al 2002). It has been estimated that representative structures for 70 percent of the total number of protein families can be modeled using known

structures (Levitt 2009) (see Section 1.5) and that by 2017, 80 percent of newly reported sequences will have a match in an existing family (Chubb et al 2010).

One can distinguish between proteins with single and multiple domain architectures (SDA and MDA). In contrast to the continuous increase of MDAs, SDAs show signs of reaching a plateau for the existing data. New protein sequences added to databases correspond mainly to already discovered protein families, and to combinations of known SDAs into MDAs (Levitt 2009). In addition, SDAs are shared by many different organisms whereas MDAs account for the diversity observed between species' proteomes (Hegyi and Gerstein 2001; Bashton and Chothia 2007; Levitt 2009).

A well-characterized property of protein families is their size distribution expressed as the number of sequences per family. The size of all known protein families follows a power law (Koonin et al 2002). As a consequence, most families are small and a large fraction of the known sequences are concentrated in a very few number of families. Figure 1.3 shows such a distribution obtained from the last release of the PFam database (PFam 25.0, Finn et al 2010).

The striking differences in the observed numbers of sequences among protein families has been a source of intense debate and speculation regarding the origin of families and their evolution (Tatusov et al 1997). There are three

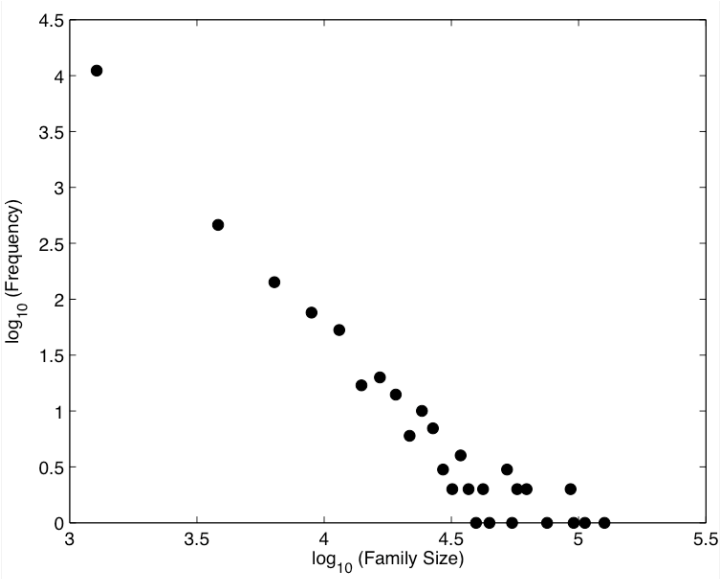


Figure 1.3. *The distribution of protein family size.* Log-log plot of the total number of sequences per protein family. Data of 12,273 families was obtained from Pfam (v25.0) (Sonnhammer et al 1998).

main, non-exclusive explanations of the differences in protein family sizes. Firstly, the designability hypothesis attributes family size differences to intrinsic structural properties of proteins. As we will discuss later, protein structures and functions differ substantially in the number of associated sequences (see Section 1.7.3).

A second, selectionist oriented view, suggests that frequent cellular functions would be overrepresented in certain protein families, and therefore sequences in those families would dominate due to the action of natural selection.

The third possible hypothesis comes from the *birth, death and innovation model* (BDIM) (Qian et al 2001; Karev et al 2002). This model proposes a mechanistic interpretation of the evolution of new protein families, similar to the process of *preferential attachment* used to explain the evolution of scale-free network architectures (Koonin et al 2002). The model assumes three fundamental underlying processes. First, protein families can be born by duplication followed by divergence (Ohno 1970). Second, they can disappear by gene loss or inactivation (Salzberg et al 2001) and third, new proteins can emerge *de novo* by shuffling and exonization (Gilbert 1978; Patty 1999), or by horizontal gene transfer (Doolittle 1999). At equilibrium the model explains the distribution of protein family size distribution well and shows that upon perturbation the model soon relaxes to a new stationary state.

Researchers' interest in the properties of protein families goes beyond the taxonomical description of the protein universe. Molecular evolution studies obtain information from the analysis of orthologous and paralogous genes, both of which correspond to subsets of genes that are part of the same protein family. The common ancestry of the sequences that belong to a protein family makes the family a key object of evolutionary studies.

1.4 Protein superfamilies

Protein families are usually grouped into *superfamilies* according to two criteria. First, sequences that belong to the same superfamily always share the same three-dimensional structure. Second, families in the same superfamily are evolutionary related. Their relationship can be claimed based on distant

sequence homology and on the conservation of their structure or function. Protein superfamilies also vary extensively in the number of families that they are composed of. Examples of large protein superfamilies include ATP-binding Rossman-like and TIM barrel structures (Orengo and Thornton 2005).

1.5 Protein structures

Protein sequences fold into three-dimensional conformations or structures. Structures are classified into hierarchical levels. The first level or *primary structure* corresponds to the amino acid sequence. The second level or *secondary structure* corresponds to three main structural patterns, namely alpha-helices, beta-sheets and turns. The third level or *tertiary structure* is defined by the three-dimensional arrangement of secondary structure elements (Branden and Tooze 1999). Depending on the composition of secondary structure elements, tertiary structures are classified as all-alpha, all-beta, alpha+beta or alpha/beta structures. Alpha/beta structures present alpha and beta elements in spatially separated regions of the same structure, whereas alpha+beta structures usually consist of interconnected patterns of secondary structure elements, for example the beta-alpha-beta pattern (Murzin et al 1995).

Analogous to the concept of sequence space, the collection of all structures is called *structural space*. The basic unit of this space is the *domain* or *fold*. A protein sequence that folds stably and autonomously into a three-dimensional conformation is known as a *structural domain* (Branden and Tooze 1999). Structural domains are usually treated as evolutionary units (Chothia et al 2003). Although the concepts of domain and fold are used interchangeably in the literature, domain is a broader term that is also used in reference to a sequence region that has some particular function or a protein family-specific sequence pattern (Hunter et al 2009). In contrast, fold has an exclusive structural connotation and refers to a coarse-grained secondary structure architecture and its connectivity. Generally, folds sit at the top of structural classification hierarchies, while domains can be defined at different places across the hierarchy (Orengo et al 1994; Murzin et al 1995). In the present document we use both concepts interchangeably, but always in reference to structure.

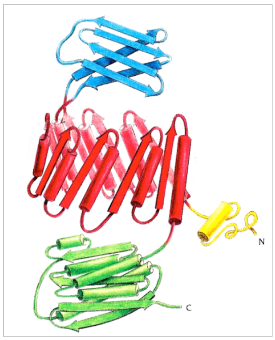


Figure 1.4. *Multiple domains in a single protein chain*. The structure shown here is that of a single-chained multidomain protein. Structural domains are represented in different colors. Reprinted with permission from Garland Publishers. (Branden and Tooze 1999). Copyright (1999).

A protein fold can be completely formed by a single protein chain or be composed of fragments that belong to different chains which come together only in the fold's spatial organization. In addition, a single chain can be composed of multiple domains. Figure 1.4 shows an example of such a case.

The identification of domains depends on the methods of structure superposition and decomposition (Orengo et al 1993; Murzin et al 1995). Today's databases have classified approximately 1,200 folds (Orengo et al 1994, Murzin et al 1995). Since the classical paper by Chothia (1992), which suggested a possible upper-limit of 800 folds, there has been sustained interest in predicting the total number of possible protein families and folds. Later predictions range between 2,000 to 10,000 folds (Zhang 1997; Wang 1998; Zhang and DeLisi 1998; Govindarajan et al 1999; Wolf et al 2000; Coulson and Moult 2002).

As the number of protein sequences in databases increases, our knowledge of protein structures also grows. Important contributors to the increase in the number of novel structures are recent international initiatives (*protein structure initiatives*, PSIs) whose aim is to characterize protein families of unknown structure and function (Dessailly et al 2009). These projects distinguish between distinct and novel structures. *Distinct* structures are those with less than 30 percent sequence identity to sequences with any known structure. *Novel* structures are those with less than 2 percent identity (Berman et al 2009). As of July 2011, PSIs have solved 2,072 distinct and 3,328 novel structures.

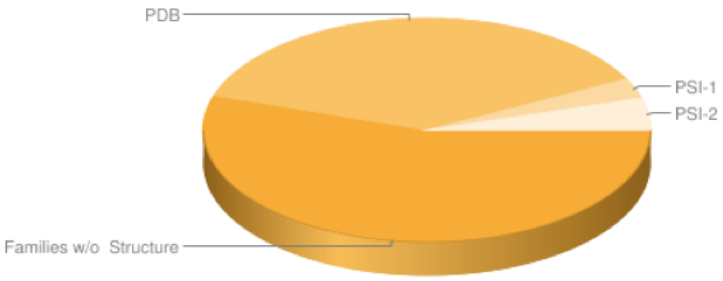


Figure 1.5. *The distribution of structures among protein families*. From the 12,273 protein families (release 25.0) (http://targetdb.sbkb.org/Metrics/MilestonesTables.html).

An important research area of PSIs is the identification of new protein families, sets of sequences that do not show any detectable homology to any other known protein family and fold. Candidate sequences are subjected to different tests that ascertain their propensity to crystallize (George and Wilson 1994). Figure 1.5 shows the fraction of protein families currently covered by the known structural space, and among those, the contribution of PSIs. Among 12,273 total protein families (PFam release 25.0; Sonnhammer et al 1998), 52 percent have at least one representative structure. Protein structure initiatives have contributed 7.5 percent to our current knowledge of protein structures (Berman et al 2009).

1.5.1 Fold complexity, stability, and kinetics.

There are multiple properties that result from the 3D structure of a protein. The *complexity of a fold* is probably the most basic property that characterizes a structure. Different measures of this complexity have been proposed (Adami 2002; Plaxco et al 1998). In general, a measure of complexity describes the number and strength of contacts that a protein structure has. One such measure is called *contact order* (CO) (Eq 1.2). It quantifies the number of non-local contacts, that is, the interactions between residues and their separation along a protein chain (Plaxco et al 1998; Baker 2000).

$$CO = \frac{1}{LN} \sum_{i,j;i>j}^{N} \Delta S_{ij} \qquad (1.2)$$

Here, L is a protein's length, N the number of contacts and ΔS_{ij} is the sequence separation along the chain in number of amino acids between residues i and j.

The number of interactions between distant residues along the chain is a reflection of the intricacy of a fold, of its complexity. A high contact order translates into a low *folding rate*, that is, the folding of a complex fold requires more time than that of a simple fold. Contact order can be used to estimate folding rates from structural information (Baker 2000).

A second measure of fold complexity is *contact density*. Contact density corresponds to the average number of contacts in a protein structure. Contact density can be estimated as the first eigenvalue of a protein's *contact map*. The contact map is a square matrix, *C*, where each entry *C(i,j)* (with $i \neq j$ and $i,j \leq L$) is equal to 1 if monomers *i* and *j* interact physically and are not adjacent on the chain. It is equal to 0 otherwise (Figure 1.6). Two amino acids interact if their distance lies below a given threshold.

Contact density correlates strongly with *thermodynamic stability* (Bloom et al 2007). The explanation is straightforward: the higher the number of contacts between two residues, the more robust is the structure to perturbations by single point mutations and by thermal noise. As a consequence, a higher number of sequences can fold into the same structure. Due to the importance of contact density as a structural determinant, it is not surprising that it also influences other protein properties, such as *evolutionary rate* (Zhou et al 2008) and *evolvability* (Bloom et al 2006; Wroe et al 2007).

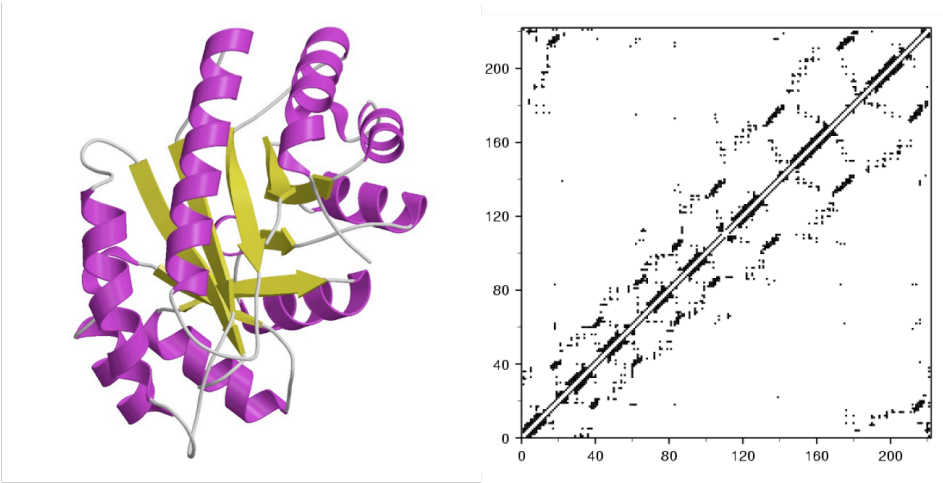


Figure 1.6. *A protein structure's contact map*. A protein structure (left) is represented as an *L*x*L* matrix where each entry, *C(i,j)* is 1 or 0 depending on the spatial proximity between residues *i* and *j* (see main text).

1.6 *Simple exact models of protein folding.*

Simple exact models (SEMs) of proteins have a long tradition in the study of protein folding and evolution (Lau and Dill 1989; Dill et al 1995; Dill and Chan 1997; Chan and Bornberg-Bauer 2002). A protein lattice consists of short polymers (L<40 monomers) composed of amino acids from a certain monomer alphabet ($\boldsymbol{A}$). These polymers fold on a grid of two or three-dimensions through discrete spatial movements. The folding of a particular polymer is dictated by an energy function encapsulated in the square matrix $U(a,b)$, whose dimension is $|\boldsymbol{A}|$, and whose entries correspond to the energy of the interaction between monomer types a and b. The total energy of a sequence folded into a particular conformation is calculated by considering only the interaction between non-adjacent contacts along the chain. The simplicity of lattice proteins allows the exhaustive enumeration of all possible sequences and their spatial conformations, and the detailed exploration of their thermodynamic properties (Lau and Dill 1989).

The quintessential lattice protein model is the HP model. In this model the alphabet is composed of hydrophobic (H) and polar (P) monomers (see Figure 1.7). The reason for this seemingly extreme simplification is the large contribution of hydrophobicity to the nature of the protein folding process (Kauzmann 1958). In the HP model only non-adjacent hydrophobic contacts contribute to the total energy. The total energy corresponds to –X, where X the number of these contacts.

```
PHHHPPPHPPHPHPHPPHHPHHHHP -10
PHPHPHPHPPHHHPHPPHPPHHHHH -13
PHPHHHPHPPHPHPHPPHHPHHHHH -13
PHPHPPPHPPHPHPHPHHHPHHHHH -12
PHPHPPPHPPHPHPHPPHPPPHHHP - 8
PHPHPHHHPPHPHPHPPHHPHHHPH -11
PHHHPHHHPPHPHPHPHHHPHHHHH -13
PHHHHHPHHPHPHPHPHHHPHHHHH -13
HHPHHHHHPPHPHPHPHHPPHHHHH -13
PHPHPHPHPPHPHHHPPHPHHHHHH -13
...
```

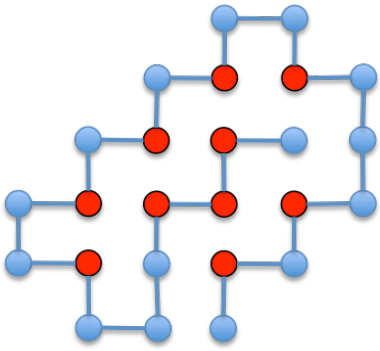


Figure 1.7. *The HP lattice model*. Left, a set of HP sequences and their corresponding energy of folding into the structure at the right. Right, lattice conformation of the sequences on the left. Hydrophobic and polar monomers are colored according to the sequence highlighted on the left.

In general, a sequence $S = \{s_1, \ldots, s_L\}$, is composed of L monomers drawn from an alphabet $\boldsymbol{A}$. The total energy E of a sequence S folded into a conformation C is computed as follows:

$$E(S,C) = \sum_{i=1}^{L} \sum_{j>i}^{L} \Delta_{ij} U(s_i, s_j), \tag{1.3}$$

where Δ_{ij} is 1 if monomers s_i and s_j, at positions i and j are non-adjacent on the chain and in contact. Δ_{ij} adopts a value of 0 otherwise. U corresponds to the energy function.

A sequence's degeneracy (g) is defined as the total number of conformations adopted by this sequence at the same minimal energy (E_m). Sequences are usually considered foldable if and only if g=1. Figure 1.7 shows an example of a set of sequences and the 2D lattice conformation into which they fold. For all these sequences, g=1.

The simplicity of lattice models allows the exploration of the effect of different monomer alphabet sizes (Buchler and Goldstein 1999a), energy functions (Wroe et al 2005), and sequence length (Irbäck and Troein, 2002) on the properties of proteins. This is why lattice proteins have been successfully applied to a diverse range of problems such as characterizing folding kinetics (Lau and Dill 1989; Sali et al 1994; Dill et al 1995), sequence determinants of folding (Abkevich et al 1994), stability (Bornberg-Bauer and Chan 1999; Wingreen et al 2004), the evolution of structure and function (Lipman and Wilbur 1991; Hirst 1999; Li et al 1996; Williams et al 2001; Li et al 2002; Blackburne and Hirst 2003), hydrophobicity (Irbäck and Sandelin 2002), energy functions (Buchler and Goldstein 1999b; Wroe et al 2005) and their reliability (Thomas and Dill 1996), reduced amino acid alphabets (Buchler and Goldstein 2000), foldability and energy landscapes (Govindarajan and Goldstein 1997; Bornberg-Bauer and Chan 1999; Xia and Levitt 2004), protein-protein interactions (Noirel and Simonson 2007), recombination (Cui et al 2002; Xia and Levitt 2002; Drummond et al 2005b; Xu et al 2005), translational accuracy (Drummond and Wilke 2008), disorder and aggregation (Giugliarellia et al 2000; Crippen and Chhajer 2002; Szilagyi et al 2008;), phenotypic mutations (Whitehead et al 2008; Drummond and Wilke 2009), as well as evolvability and adaptation (Wroe et al 2007; Chen et al 2010; Bornberg-Bauer et al 2010).

1.7 The protein sequence–structure relationship as a genotype-phenotype map

Genotypes encode the information required to produce organismal traits, which are commonly called *phenotypes*. Although we usually think of genotypes as DNA, there are different levels of complexity at which genotypes can also be represented. Those levels may be simply understood as coarse-grained versions of the genetic information in DNA. For instance, we may say that a set of chemical reactions constitutes the genotype of a *metabolic network*, whose phenotype comprises the molecules that the network synthesizes (Matias Rodrigues and Wager 2009). Similar representations of the concepts of genotype and phenotype have been used to study the sequence-structure relationship in RNA and proteins (Lipman and Wilbur 1991; Schuster et al 1994), in gene regulatory circuits (Wagner 1996), and even in man-made systems such as hardware-circuitry (Raman and Wagner 2011) and genetic programs (Altenberg 1994) (see Table 1.1).

One can think of a function ***F*** that maps each possible genotype from a set *G* to a particular phenotype in the set *P* of all possible phenotypes (***F***: *G*->*P*). This function has been called a *genotype-phenotype map* (Alberch 1991). In the case of RNA and proteins, the set of genotypes (*G*) arises from combinations of elements of a genetic alphabet (***A***), that is from sequences; whereas the set of phenotypes (*P*) arises from the spatial conformations that these sequences form. *G* and *P* can also be viewed as spaces of protein sequences and structures.

Some of the most important questions in evolutionary biology regard how the information encoded in the genotype produces a phenotype. The genotype-phenotype map describes the variability of phenotypes, that is, their potential to change in response to genotypic change, a property that has also been called *evolvability* (Dawkins 1989). Furthermore, the genotype-phenotype map relates to a diverse and interrelated range of phenomena, such as *genetic robustness* (de Visser et al 2003), *adaptation* (Wagner and Altenberg 1996), and *development* (Waddington 1952).

Despite their complexity, the genotype-phenotype maps of RNA and proteins are the simplest natural genotype-phenotype maps. There are several reasons why genotype-phenotype maps of macromolecules deserve attention. First, they have small genetic alphabets (4 types of nucleotides in the case of RNA

Table 1.1. *Examples of genotype-phenotype maps.*

System	*Genotype*	*Alphabet*	*Phenotype*
RNA	Sequence	Nucleotides	Structure
Proteins	Sequence	Amino acids	Structure
Metabolic networks	Set of reactions	Reactions	Metabolites
Regulatory circuits	Gene interactions	Genes	Expression pattern
Language	Sentence	Words	Meaning
Hardware	Logic circuit	Circuit	Logic function
Genetic algorithms	Set of operations	Operation	Performance

and amino acid alphabet sizes that range between 2 and 20 in the case of proteins). Second, their phenotypes are the product of physicochemical interactions codified in the genotype. Third, for about 50 years, structural biology has studied the physicochemical basis of sequences and structures. These efforts resulted in millions of sequences and thousands of structures. Although attention has focused predominantly on proteins, more recently RNA has also attracted immense interest. In addition, research on polymer physics provided simple folding models that allow the exploration of questions that one can currently not address for natural macromolecules. Also, algorithms exist that perform *in silico* folding of RNA secondary structure (Zuker and Stiegler 1981; Hofacker et al 1994). I next review our current knowledge of the protein genotype-phenotype map.

1.7.1 The Chothia-Lesk plot

The first general assessment of the relationship between protein sequences and structures was carried out in 1986. Chothia and Lesk (Chothia and Lesk 1986) compared 32 pairs of homologous proteins with known structures. They aligned their structures and compared their identity to sequence identity in the superimposed regions. Figure 1.8 shows a version of their analysis based on recent data. Similar comparisons reached the same qualitative observations (Sander and Schneider 1991; Orengo et al 2001; Reeves et al 2006). I next discuss some of these observations. First, pairs of proteins that have between 30 and 100 percent sequence identity show considerable structural identity. At structural identities higher than 80 percent (Figure 1.8, green dashed line), two proteins are usually considered to have the same fold

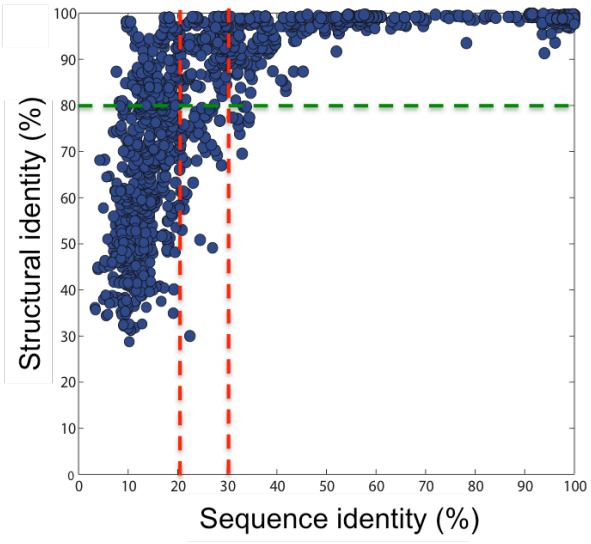


Figure 1.8. *The Chothia-Lesk plot*. The figure shows sequence versus structural identity of 10^4 pairs of protein structures. The plot shows sequence identity (in the superimposed region) versus structural identity, measured by a method of vectorial representation of structures (Ortiz et al 2002).

(Orengo and Taylor 1990). This finding had important consequences for the development of structure modeling methods. It shows that the structure of a sequence A can be approximated by any known structure B whenever the alignment between sequences A and B shows more than 30 percent of sequence identity. The higher the sequence identity, the higher the confidence of obtaining an accurate model of structure A. This approach is known as *comparative modeling* (Sali and Blundell 1993; Mati-Renom et al 2000) and today represents the most successful method to model protein structures (Marti-Renom et al 2000).

Secondly, in the region between 20 to 30 percent sequence identity (Figure 1.8, red dashed lines), also called the *twilight zone*, many pairs of proteins show a drastic reduction in structural conservation, although some of them still have similar folds. This means that once the sequence identity of two proteins falls into the twilight zone it is not possible to assume any longer that the two sequences share the same structure. From an evolutionary perspective, the twilight zone implies that no conclusions related to the common ancestry of two sequences can be assumed. Structural studies aiming to predict side-chain coordinates and packing of residues in proteins have revealed some of the structural changes that occur in the twilight zone. At 50 percent sequence

identity, an average backbone RMS error (Root Mean Square error, a measure of structural deviation, the converse of identity in Figure 1.8) of 1 Ångstrom between the superimposed structures translates into 65 percent of side-chain prediction accuracy. However, when the twilight zone is reached, a backbone RMS values of 1.9-2.0 Ångstroms translates into only 29 percent of correctly predicted side chains. In other words, in the twilight zone, backbone conservation is unable to constrain the correct packing of residue side chains, which leads to major structural deviations (Chung and Subbiah 1996).

A third general observation is that structures are more conserved than sequences during evolution (Ptitsyn and Volkenstein 1986).

1.7.2 Insights from simple exact models: neutral networks in sequence space.

Simple exact models have been essential in exploring the organization of the sequence-structure map. To be able to discuss the contribution of these models to our current knowledge of the protein genotype–phenotype map, I need to briefly provide further information on protein sequence space.

I have already defined sequence space as the collection of all possible protein sequences (Section 1.2). One of the main advantages of formalizing the space of protein sequences in terms of an explicit geometric object is that it provides a natural distance metric and mathematical graph formalisms.

Formally, the sequence space of proteins, $\mathcal{S}(L,\mathbf{A})$, is defined by the sequence length (L) and the amino acid alphabet $\mathbf{A}$ ($2 \leq |\boldsymbol{A}| \leq 20$). The sequence space comprises $|\boldsymbol{A}|^L$ sequences. The *dimension* (n) of the sequence space is defined by: $n = L(|\boldsymbol{A}|-1)$ and in the case of the simplest amino acid alphabet, $|\boldsymbol{A}|=2$, n is simply equal to L. Here, I precise an important distinction between *protein sequence space* ($\mathcal{S}$), composed of all possible sequence combinations, and the *protein universe* ($\mathcal{U}$), which corresponds to those sequences present in nature (Ladunga 1992; Levitt 2009).

Protein sequence space can be described as a *generalized hypercube graph* Q_A^L (Reydis et al. 1997). For $|\boldsymbol{A}|=2$, this graph is equivalent to a *hypercube,* a generalization of a three-dimensional cube to n dimensions also called n-cube. A *hypercube graph* Q_2^L, or L-cube (because for $|\boldsymbol{A}|=2$, $n=L$), is a graph whose vertices are the vertices of this hypercube; an edge connects two vertices in this

graph if the corresponding vertices of the hypercube are adjacent. Figure 1.9 shows the construction of L-cubes of increasing dimensions. Due to their peculiar properties hypercubes have been applied to many areas of inquiry, such as parallel computing (Seitz 1985), coding theory (Aiello and Leighton 1991), as well as fitness landscapes and speciation (Wright 1932; Gavrilets 2004).

The L-dimensional hypercube graph Q_2^L = ($\boldsymbol{V}$,$\boldsymbol{E}$) has 2^L nodes ($\boldsymbol{V}$) and $L2^{L-1}$ edges ($\boldsymbol{E}$). It is said that a hypercube has order p = |$\boldsymbol{V}$| and size q = |$\boldsymbol{E}$| (Morgan 1989). Each node of Q_2^L has exactly L neighbors. A *distance d(x,y)* between nodes x and y is formally defined over the metric space $\mathcal{S}(L,\mathbf{A})$ = Q_2^L, as the number of bits (or in general, symbols) that differ between the binary labels of those two nodes. This metric is also called the Hamming distance (Hamming 1958). In other words, the Hamming distance between two sequences, $d(s_1,s_2)$ in sequence space $\mathcal{S}$, corresponds to the minimum number of amino acid changes needed to transform sequence s_1 into s_2. A metric space can also be viewed as a *shape space,* a concept introduced originally by Perelson and Oster (1979).

The *diameter* of the graph $\boldsymbol{G}$, $D(\boldsymbol{G})$ is the maximum distance between any pair of nodes. Therefore, $D(Q_2^L) = L$.

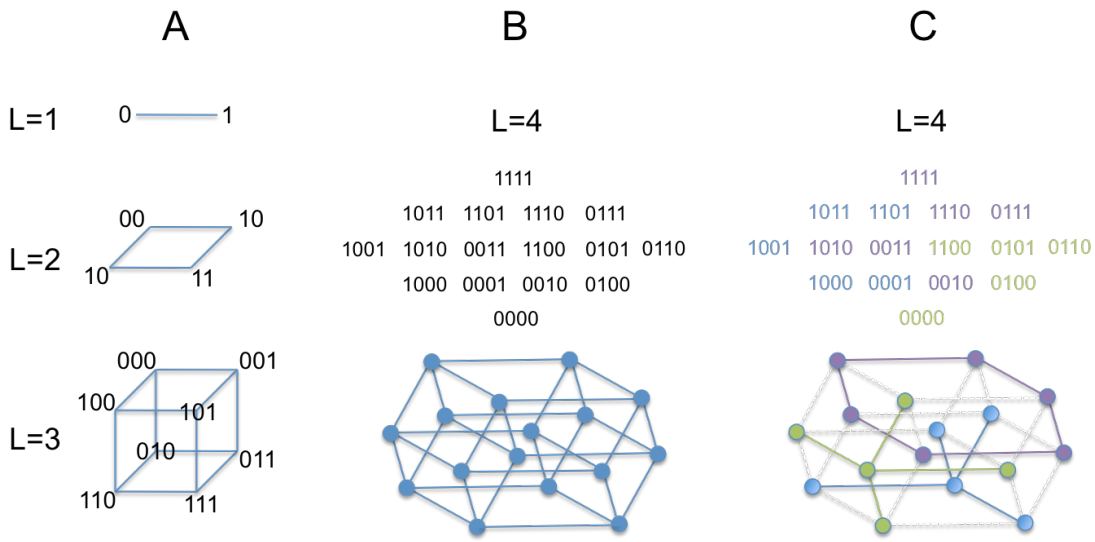


Figure 1.9. *The construction of a hypercube*. A. Progressive construction of the 3-cube by increasing the length of a binary sequence. B. The 2^4=16 vertices of the four-dimensional hypercube graph Q_2^4 represented as binary strings of length four where $\boldsymbol{A}$={0,1}. Below the bit strings is a geometric representation of the graph. C. A partition of the hypercube shown in panel B into 3 colored subspaces.

The *n-dimensional ball* (or *k-ball*), $B_k^n(x)$, centered at node x and of radius $k \geq 0$, consists of all nodes whose distance to x is no higher than k. The smallest value of k of $\underline{B_k^n(x)}$ that contains all edges in $\boldsymbol{E}$, is defined as the radius of the graph $\boldsymbol{G}$, R($\boldsymbol{G}$). Therefore, R(Q_2^L) = L. Note that, D(Q_2^L) = R(Q_2^L).

A *generalized hypercube graph* ($Q_{|A|}^L$) admits more than 2 symbols for each string that represents a vertex ($|\boldsymbol{A}|>2$). Exactly as in a hypercube graph, two vertices are connected if their strings differ in exactly one symbol. However, the geometric interpretation is less intuitive. In the case of proteins, an alphabet size of $|\boldsymbol{A}|=2$ could represent, for example, the HP (hydrophobic-polar) model, where the 20 amino acids are classified according to their polarity into hydrophobic (H) or polar (P). An alphabet size of 20 would comprise the full set of 20 natural amino acids. In the following, the protein sequence space is defined as a generalized hypercube ($\mathcal{S} = Q_{|A|}^L$).

The first insights into protein genotype-phenotype map organization came from a study carried out in 1991 by Lipman and Wilbur (Lipman and Wilbur 1991). This study used the HP model to test an earlier conjecture by John Maynard-Smith on the conditions for the evolution of protein functions (Maynard-Smith 1970). Maynard-Smith argued that protein functions can evolve if and only if, $f(|\boldsymbol{A}|-1)L > 1$. Here, as described above, $|\boldsymbol{A}|$ and L correspond to amino acid alphabet size and protein length respectively; while f corresponds to the fraction of functional neighbors that are as active as the sequence under study. Lipman and Wilbur used two-dimensional HP lattice models that were 16 and 19 amino acids long. They showed that sets of sequences that fold into the same conformation extend over large regions of sequence space, and form mutational or *neutral networks* (Lipman and Wilbur 1991). Figure 1.10 shows the largest neutral network observed in the HP model of length 25. The network is formed by sequences folding into the conformation shown in Figure 1.10A. The color code in Figure 1.10B represents stability of each sequence in the neutral network (see Section 1.7.4).

Conclusions from the Wilbur and Lipmann study are supported by subsequent work. For instance, studies using energy functions derived from

atomic interactions of protein crystal structures and *in silico* point mutations, suggest that far-reaching mutational networks also exist for natural proteins (Babjide et al 1997; Babajide et al 2001).

1.7.3 Insights from simple exact models: the designability hypothesis.

A second major contribution from lattice models was the observation of heterogeneous distribution of the number of sequences per structure.

The first evidence of a fold's tendency to vary in terms of the number of associated sequences came from fold classification systems (Orengo et al 1994). Some compiled folds were extensively represented in natural proteins, whereas most folds had only one associated sequence. The common folds were called *superfolds* (Figure 1.2 shows nine of them first identified in 1994 (Orengo et al 1994)). Superfolds have peculiar physicochemical properties. They possess relatively simple three-dimensional structures (Orengo et al 1993; 1994), are symmetric (Hartling and Kim 2008), particularly stable (Bloom et al 2005) and fast folders (Dias and Grant 2006; Ferrada and Wagner 2008).

Although the existence of superfolds may be the result of a biased sample of structures in nature, the use of lattice models revealed a similar pattern of the distribution of the number of sequences per fold. Exhaustive enumerations of

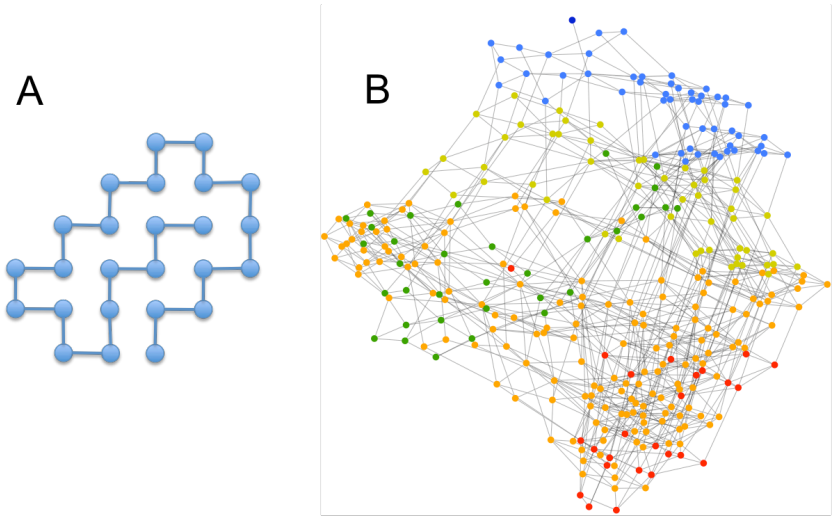


Figure 1.10. *The largest neutral network of the two-dimensional HP lattice model of length 25*. A. The conformation into which sequences of this neutral network fold. B. The neutral network is composed of 326 sequences and 942 transitions between single point mutants. The color code represents the stability of each sequence. It ranges from -8 contacts (blue, top) to -14 contacts (red, bottom).

protein lattice sequences showed that a high fraction of the foldable set of sequences adopts only few conformations, while most conformations are associated with single sequences (Li et al 1996). Based on this observation, common conformations were (misleadingly) called *designable* and showed properties that echoed the stability, folding kinetics and symmetry properties seen in natural proteins (Helling et al 2001). Thus, despite their simplicity, protein lattices showed that the heterogeneous distribution of sequences versus structures may be a physicochemical, inherent feature of proteins, rather than a result of sampling bias or the effects of natural selection on protein folds (Li et al 1996; Melin et al 1999).

The strong support given by protein lattices to the heterogeneous distribution of the number of natural sequences per structure led to the so called *designability hypothesis* (Li et al 1996). This hypothesis states that a small fraction of folds are associated with a high number of compatible sequences (Kussell 2005).

Protein designability correlates strongly with contact density, the average number of contacts per amino acid (England and Shakhnovich 2003). This observation has allowed the theoretical study of protein designability and its relation to other properties of proteins (Bloom et al 2005; Ferrada and Wagner 2008). Recent studies have concentrated on the evolutionary consequences of designable folds, showing that designability may affect a protein's evolutionary rate (Zhou et al 2008). Through its relation to protein stability, designability may promote *evolvability*, the propensity of a phenotype to produce new phenotypes (Bloom et al 2006).

1.7.4 Insights from simple exact models: the energy landscape and marginal stability.

Subsequent studies shed further light on the protein GP map organization by explicitly exploring the distribution and organization of sequences that fold into the same conformation (Bornberg-Bauer 1997; Bornberg-Bauer and Chan 1999). Using the HP lattice model of sequences with L=18, these studies showed that taking into account sequence stability, neutral networks have a *funnel-like* organization. In this organization, more stable proteins are at the center of the

network and sequences increase in energy proportionally to their distance from the center (Figure 1.11). The sequence in the center has been called *prototype sequence* (Bornberg-Bauer 1997; Bornberg-Bauer and Chan 1999). Additionally, neutral networks seem to be isolated from one another in sequence space (Bornberg-Bauer and Chan 1999). The isolation of protein neutral networks space has been highlighted in other lattice studies (Yahyanejad et al 2003) and conjectured for natural proteins (Nishikawa 1993). However, the extent to which this phenomenon depends on properties of sequence space and occurs in natural proteins is not currently clear.

In summary, observations from natural proteins and simple exact models suggest four main properties of the protein genotype-phenotype map. First, there are more sequences than structures. Second, the number of sequence per

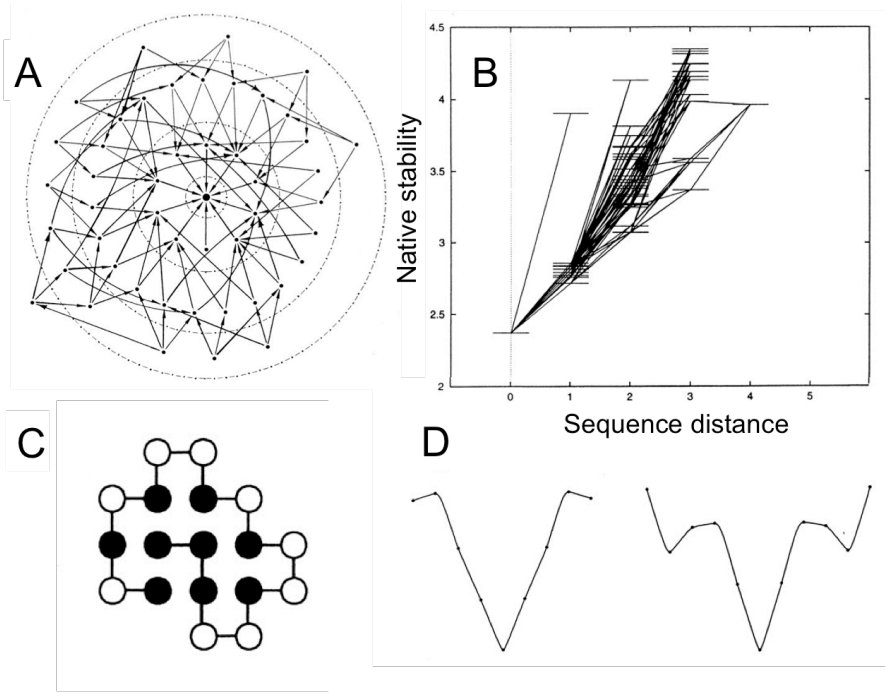


Figure 1.11. *The organization of protein neutral networks in sequence space*. A. Neutral network of a set of sequences connected by single point mutations and folding into a unique conformation (represented in C). Sequences are represented as dots. Consecutive circles in dashed lines represent distances between single point mutant sequences. Arrows correspond to favorable transitions toward more stable sequences. The center corresponds to the prototype sequence. B. Sequence stability versus distance from the prototype sequence. C. Conformation into which all sequences represented in A fold. Black dots represent hydrophobic monomers, white represent polar monomers. D. Two possible models of the funnel. Figure modified from figures 1 and 2 in (Bornberg-Bauer and Chan 1999). Reproduced with permission from Proc Natl Acad Sci USA, copyright (1999).

structure has a non-uniform distribution. A large fraction of sequence space folds into a small set of conformations, while the vast majority of conformations are only realized by a single sequence. Third, sequences form neutral networks that cover large regions of sequence space. Fourth, although some of these networks can be connected to each other by few point mutations, the majority may inhabit isolated regions of the space.

1.8 Protein functions.

The ultimate goal of protein science is to describe protein function. Only through the functional description of a macromolecule, its biology can be understood. Structure and function relate to each other in a non-trivial manner. The challenge is to infer protein functions from sequence, structure, and from systems biology data.

1.8.1 Functional annotation

One of the most basic problems in protein science is to precisely define protein *function*. Most of the functional annotation errors of proteomes stem from semantic properties of the function concept. For instance, it has been estimated that the functional annotation of the human genome using different methods yields 20 to 30 percent of disagreement about the functions of proteins (Raes et al 2007).

Annotation means inference of function by means of sequence and structure information. This process is in essence similar to comparative modeling of protein structures and because of that, it is intrinsically connected to the properties of protein sequence space. If we want to know the function of sequence A, we proceed by searching for the most similar sequence B with known function. In the best case, sequence B has a function known from experiment.

According to their reliability, one can distinguish between different levels of functional annotation. First, the most reliable information about a protein's function is obtained by experiment. Data from Uniprot, the most comprehensive resource of proteins, indicates that we possess this information for

approximately one percent of the total protein sequences known today (The Uniprot Consortium 2010).

A second source of functional information is sequence homology. Evolutionary information readily assigns functions to orthologous and paralogous proteins. Although it was shown that, as in the case of structures, 40 percent of sequence identity usually provided reliable functional annotations (Devos and Valencia 2000), it was later observed that proteins with different functions may have high sequence identity (Rost 2002). For example, even at 70 percent of sequence identity, around 10 percent of protein pairs perform different functions (Rost 2002).

A third possibility to annotate protein functions is through sequence patterns of essential protein residues. Such patterns are identified using entropy measures of residue site conservation (Karchin et al 2005), regular expressions (Apweiler et al 2000), graph theoretical tools (Ruepp et al 2004), and phylogenetic evidence (McAuliffe et al 2004). This type of classification is particularly useful for enzymes and ligand binding proteins, where conservation of a small set of residues or *motif* can be diagnostic of conserved function (Apweiler et al 2000).

A fourth approach integrates different information sources, such as those I just described with systems biology data into automatic pipelines of functional annotation (Erdin et al 2011).

Functional annotation by any of the methods mentioned above can be applied to around 65 percent of the total number of sequences known today. 35 percent of these sequences still have '*putative*', '*uncharacterized*' or '*hypothetical*' functions (Raes et al 2007). It is currently possible to annotate 96 percent of the *E. coli* genome, and the function of 73 percent of gene products in an average genome.

One main difficulty in the annotation of functions is that homology is not always synonymous with functional conservation. Gene duplication, for example, may create proteins that eventually perform very different functions (Conant and Wolfe 2008). An additional phenomenon that complicates annotations is *functional promiscuity* and *multifunctionality*.

The concept of functional promiscuity has been used in reference to a wide variety of related phenomena (Jensen 1976; Khersonsky and Tawfik 2010). I here refer to it as the ability of a protein enzyme to catalyze a reaction whose substrate is different from the one the enzyme has evolved for (Khersonsky et al 2006). We will discuss later in more detail the role of enzyme promiscuity in the evolution of protein functions.

Enzyme promiscuity is a special case of a more general phenomenon called multifunctionality (Jeffery 1999). Multifunctional proteins are found to perform different functions under physiological conditions. Protein multifunctionality reveals additional complexities of the concept of function. However, despite being problematic in the process of functional annotation, multifunctionality may be important for the evolution of new functions.

1.8.2 Enzymes

Among protein functions those of enzymes are the best annotated. Two main factors have contributed to this fact. First, enzymes were early recognized as important molecules in cellular physiology. Second, enzymes are amenable to study by tools from chemistry and kinetics. These two factors and detailed structural knowledge prompted the early classification of enzymes.

The first enzyme classification was developed during the early 1960s by the *International Union of Biochemistry and Molecular Biology* (IUBMB) and is known as the *Enzyme Commission* classification (EC) (Bairoch 2000). Enzymes are classified based on a four level hierarchy. The top level comprises six enzyme classes, namely *oxidoreductases, transferases, hydrolases, lyases, isomerases* and *ligases*. Each class is subdivided into three further hierarchical levels whose interpretation differs among classes. In this classification system, individual enzymes are assigned a four-digit number where each digit reveals increasing details about enzyme function. For example, the enzyme *tryptophan synthase* with EC number 4.2.1.20 is a lyase that catalyzes the conversion of *indole* and serine to *tryptophan*. Although the EC classification has well-known limitations (eg. see Todd et al 2001), it is the best-established and most widely used system for classifying enzymes, which are the most prominent protein class. (By March 2010, 57 percent of proteins in the Protein Data Bank - Berman et al 2000 - had

at least one enzymatic function). Today the EC classification contains around 4,000 different enzymes (Bairoch 2000).

1.8.3 Marginal stability and protein functions

It has been repeatedly observed that proteins are marginally stable, with free energies that fluctuate between -5 to -10 kcal/mol. This means that observed native structures can be transformed through point mutations to more stabilized versions of the same fold. There are reasons to expect that nature would favor more stable proteins, among them that high stability translates into resistance to denaturation (Wagner and Wuthrich 1979), aggregation (Lomas and Carrell 2002), proteolysis (Hubbard et al 1994), and evolvability (Bloom et al 2006). There are fundamentally two hypotheses aiming to explain the origin of marginal stability in proteins.

Some studies suggest that protein have evolved toward marginal stability. The arguments rely on common features of functional proteins such as the presence of flexible regions that promote ligand binding or conformational change (Namba 2001). In addition, destabilizing mutations usually participate in the acquisition of new functions and, relatedly, amino acids substitutions in enzyme active sites usually contribute to reduce protein stability (Wang et al 2002; Chen et al 2005). These observations have been expressed in the context of a tradeoff between stability and functionality (Bloom et al 2004). A recent study by Tokuriki et al (Tokuriki et al 2008) identified protein mutations of laboratory evolution experiments and observed that changes in stability induced by mutations that confer new functions, were as destabilizing as the average destabilizing effect of any other mutation in the same protein. In addition, they found that adaptation of proteins to new functions might be fostered by *compensatory mutations* in other regions of the same protein (Tokuriki et al 2008; Tokuriki and Tawfik 2009a).

In contrast to these adaptationist-oriented explanations, a second view on marginal stability, championed by Richard Goldstein and collaborators, suggests that marginal stability arises as a result of neutral evolution (Taverna and Goldstein 2002; Williams et al 2006; Goldstein 2010). This view predicts the origin of marginal stability as a consequence of mutation – selection balance in a

high dimensional sequence space. As mentioned previously (see section 1.7.4), one of the features of protein neutral networks is their organization into a funnel-like energy landscape in sequence space. According to this perspective, the consensus or prototype sequence is the most stable folder, and stability decreases proportionally to deviations from the consensus. The high dimensionality of sequence space implies that the number of sequences around the prototype grows exponentially with divergence. Therefore, by mutation – selection balance sequences are much more likely to inhabit regions of the network with submaximal stability.

1.9 Plan of the dissertation

This dissertation is composed of five more chapters. The next chapter is entitled "A comparison of genotype-phenotype maps for RNA and proteins" (Ferrada and Wagner, submitted). In this chapter, I compared properties of simple models of proteins and RNA. I find that many fewer proteins than RNA molecules fold, but they fold into many more structures than RNA. In consequence, protein phenotypes have smaller genotype networks whose member genotypes tend to be more similar than for RNA phenotypes. Neighborhoods in sequence space of a given radius around an RNA molecule contain more novel structures than for protein molecules. I compare this property to evidence from natural RNA and protein molecules, and conclude that RNA genotype space may be more conducive to the evolution of new structure phenotypes.

In the third chapter, "Protein robustness promotes evolutionary innovations on large evolutionary time scales" (Ferrada and Wagner 2008); I explore the interplay between structural robustness and functional innovations. I study protein domains conserved through all extant organisms and show that more robust proteins have a greater capacity to produce functional innovations.

The fourth chapter is entitled "Evolutionary innovations and the organization of protein functions in genotype space" (Ferrada and Wagner 2010). In this study I show that different neighborhoods of genotype space contain proteins with very different functions. This property both facilitates evolutionary innovation through exploration of a genotype network, and it

constrains the evolution of novel phenotypes. I show that the space of protein functions is not homogeneous, and different genotype neighborhoods tend to contain a different spectrum of functions, whose diversity increases with increasing distance of these neighborhoods in sequence space.

In the fifth chapter, I use mathematical formalisms introduced in Section 1.7.2, to explore the organization of protein genotype space. I show that neutral networks' consensus sequences distribute randomly in sequence space at distances that scale according to the amino acid alphabet size. Next I show how these observations may be used to deepen our understanding of the evolution of the protein genotype-phenotype map. I finally propose a simple model that aims to reconcile the extensive size variation observed in natural proteins with the genotype space concept.

The sixth and last chapter concludes by summarizing some of the main observations made in this dissertation.

2. A comparison of genotype-phenotype maps for RNA and proteins

Ferrada E and Wagner A. A comparison of genotype-phenotype maps for RNA and proteins. (*Submitted*).

A comparison of genotype-phenotype maps for RNA and proteins

Evandro Ferrada[1,3] and Andreas Wagner[1,2,3]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, CH-8057 Zurich, Switzerland.

[2]The Santa Fe Institute, Santa Fe, New Mexico 87501, USA.

[3]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland.

Abstract

The relationship between the genotype (sequence) and the phenotype (structure) of macromolecules affects their ability to evolve new structures and functions. We here compare the genotype space organization of proteins and RNA molecules to identify differences that may affect this ability. To this end, we computationally study the genotype-phenotype relationship for short RNA and lattice proteins of a reduced monomer alphabet size, to make exhaustive analysis and direct comparison of their genotype spaces feasible. We find that many fewer protein molecules than RNA molecules fold, but they fold into many more structures than RNA. In consequence, protein phenotypes have smaller genotype networks whose member genotypes tend to be more similar than for RNA phenotypes. Neighborhoods in sequence space of a given radius around an RNA molecule contain more novel structures than for protein molecules. We compare this property to evidence from natural RNA and protein molecules, and conclude that RNA genotype space may be more conducive to the evolution of new structure phenotypes.

Protein and RNA perform myriad structural, regulatory, and enzymatic functions inside organisms. RNA may have played a more important role early in life's evolution, but this role has since been usurped by proteins, especially in catalysis (1). Both protein and RNA molecules have a primary structure, their amino acid and nucleotide sequence. This sequence can form secondary structures that comprise alpha-helices and pleated beta-sheets for proteins, and planar folds that arise through intramolecular pairing of complementary bases for RNA. The secondary structure is the basis for a molecule's tertiary structure, that is, the three-dimensional fold required for many molecular functions, such as enzymatic catalysis.

To understand how the many known functions of protein and RNA arose in evolution, and to understand how new functions originate, it is insufficient to study individual molecules. One must study the collection of all known molecules in the context of an organizing framework. This framework is provided by the concept of sequence space or *genotype space* (2), the collection of all possible nucleotide or amino acid sequences. Specifically, because evolution is driven by genotypic change, one must study how changes in a molecule's genotype affect its phenotype, that is, its fold and its function. Many past efforts aimed at characterizing this genotype-phenotype relationship. In proteins, these efforts rely both on experimental data from known proteins, and on simple models of protein folding, such as lattice proteins (3-5). In RNA, where many fewer tertiary structures are known than for proteins, such efforts have until recently (6) largely focused on RNA secondary structures (7-9). A limitation of this focus is that secondary structures are necessary but not sufficient for the function of many RNA molecules.

Existing work indicates an important similarity between protein and RNA genotype spaces: molecules with the same structure can have widely divergent sequences (10-11); and these molecules can typically be connected by a series of single amino acid or nucleotide changes that leave the structure unchanged (12-13). In other words, molecules with the same structure form large connected networks in genotype space. These networks are variously called mutational networks, neutral networks (9, 14), or genotype networks. For proteins, pertinent evidence comes from phylogenetic analysis of many proteins

with known structure and functions. An example is globin molecules, which are oxygen-binding globular proteins that probably have a common evolutionary origin (15). Throughout this evolutionary history, globins have preserved a common structure and biochemical activity, despite having diverged to a great extent in their sequences: Only 12 percent of amino acids are preserved among known globins (10). In RNA molecules, most pertinent evidence about the organization of genotype space comes from computational predictions of RNA secondary structure (9). For example, RNA molecules that adopt the cloverleaf secondary structure characteristic of transfer RNAs may differ in more than 90 percent of their nucleotides (11, 16).

The purpose of this contribution is to compare the relationship between sequence and structure for proteins and RNA. To be able to study this relationship systematically, we mostly use simple models of structure formation. Specifically, we study short proteins with a simplified amino acid alphabet and their fold on a two-dimensional lattice, as well as short RNA molecules of a simplified nucleotide alphabet and their planar, secondary structure fold. Such models (3, 8) allow one to explore the genotype-phenotype relationship of macromolecules for tens of thousands of genotypes and phenotypes. These models are relevant for understanding larger and more complex biological molecules, as shown by thermodynamical calculations, folding studies, and evolutionary studies (17-19).

With these models, we explore general folding statistics, and the organization of phenotypes into connected networks in sequence space. Most importantly, while similar work has been carried out for RNA and proteins separately (4, 9, 20, 21), our main purpose is to juxtapose and compare RNA and proteins in this regard. We supplement this comparative analysis with a limited analysis of recent empirical data from natural protein and RNA structures.

Results

The sequence space of RNA secondary structure and protein lattice models

An intrinsic problem of comparing RNA and protein sequence spaces is that they possess different dimensions. We here alleviate this problem by studying sequences with a reduced alphabet size A, that is, a reduced number of

different monomers that can occur in a molecule. In general, the dimension of sequence space is given by $L(A\text{-}1)$, where L corresponds to the length of a sequence, and A to the size of the monomer alphabet (A=4 and A=20 for biological RNA and protein molecules). In this work, we consider RNA and protein sequence spaces of dimension 25. Specifically, we analyze model proteins of length L=25 that consist only of two types of amino acids, hydrophobic (H) and hydrophilic (P for polar). In other words, we use the well-studied HP model of protein folding (3) whose alphabet size is equal to A=2. For RNA, we use molecules of length L=25, and a reduced alphabet size of two instead of four (A,U,G,C) nucleotides. Specifically, we consider sequences composed only of G and C nucleotides (see the supplementary material for a discussion of the AU alphabet). We compute the fold of HP model proteins on a 5x5 protein off-lattice using standard methods (22), and we compute the minimum free energy (mfe) secondary structures of RNA sequences using the Vienna RNA package (8; see also Methods). Below, we refer to the two data sets that emerge from these computations as HP25 and GC25.

We are well aware that many researchers have studied RNA and protein sequence spaces individually (9, 14, 20-21). However, none of these studies have directly compared sequence-structure relations of proteins and RNA, which is the main purpose of our analyses.

The total number of possible protein or RNA sequences in our model system is 2^{25}. We first analyzed which of these sequences fold into a unique structure. In the protein (HP25) data set, only 2 percent of sequences do (Table 1). We call such sequences *uniquely foldable*. In the RNA (GC25) data set 99.9 percent of sequences fold into unique secondary structures (Table 1). These statistics and the numbers of different structures that these sequences form are summarized in Table 1.

The observation that there are many fewer foldable proteins than RNA molecules is perhaps the most prominent difference between proteins and RNA. This observation is consistent with experimental evidence from biological molecules. For example, soluble and compact protein structures are rare in random protein libraries (23-24), whereas RNA molecules taken from a random

library possess a high probability of collapse into compact and ordered structures (25). The ultimate causes of these differences lie in the chemistry and folding mechanisms of RNA and proteins (26). However, more than in their causes, we are here interested in the consequences of these differences.

The distribution of sequences versus structures

Tables 1 and 2 show that in both the protein and RNA data sets, many more foldable sequences than structures exist. This implies that any one structure is typically formed by multiple uniquely foldable sequences. Figure 1 shows the distribution of the number of sequences per structure for both proteins and RNA. The figure shows that the number of sequences per structure is highly heterogeneous and varies over several orders of magnitude for both proteins and RNA. Taken together, the sequences that form those structures with many associated sequences account for a majority of foldable sequences. For example, the structures whose associated number of sequences is in the top 10 percent (among all structures) account for 85 percent of foldable sequences in the case of RNA, and for 47 percent in the case of proteins. This property has been observed separately for both RNA (9) and proteins (27).

Figure 1 also shows another important difference between RNA and proteins: except for those structures that are formed by the smallest number of sequences, protein structures are generally formed by fewer sequences than RNA structures. This is evident from the much steeper slope of the protein data in Figure 1. It is a consequence of the fewer uniquely foldable sequences and the higher number of structures for proteins (Table 1) (25, 28).

Neutral networks in sequence space

We next analyzed how different the sequences are that fold into any one structure. To this end, we first define a genotype set (or neutral set) as the collection of all sequences that fold into a given structure. We define a genotype network (or neutral network) as a collection of sequences that fold into the same structure and that can be connected to each other through a sequence of single monomer changes, none of which change the structure (9, 20). A single genotype set can contain one or more genotype networks.

Table 2 summarizes observations from this analysis. The first notable feature is that proteins form many more structures than RNA, and thus have many more neutral sets, but each such set has many fewer sequences (7.1 sequences per structure for HP25 proteins, versus more than 1,000 sequences per structure for GC25 RNA, Table 2). The latter observation is another consequence of the fact that fewer proteins are uniquely foldable. The largest genotype set comprises 326 sequences for HP25 proteins, but 202,217 sequences for GC25 RNA molecules. In both proteins and RNA, however, the vast majority of genotype sets is small (Figure S1 and Figure 1).

We next asked in how many monomers sequences within a genotype set typically differ. Figure 2 shows the distributions of this average sequence distance. HP25 protein sequences with the same structure are typically much more similar to each other (mean ± std. deviation: 1.3±1.1 monomer differences) than GC25 RNA sequences (7.4±3.3 differences). *Maximum* distances between sequences with the same structure are also much smaller in proteins than in RNA (mean ± std. dev. 2.9±2.7 and 15.9±8.9 monomer differences, respectively; Figures 2B and 2D). While small maximal sequence distances dominate for proteins (50 percent of the HP25 genotype sets show maximum distances shorter than 3 point mutations), this is not the case for RNA (50 percent of the GC25 genotype sets have maximum distances larger than 18-point mutations). Moreover, 32 percent of RNA genotype sets have a maximum distance of 25 and thus extend all the way through genotype space, but none do so for the HP25 proteins.

As mentioned above, genotype sets may be composed of more than one connected component or genotype network. The number of genotype networks per genotype set is smaller in proteins than in RNA (Table 2), and genotype networks contain on average fewer sequences for proteins than for RNA. Figure S2 shows the distribution of mean and maximum distances between sequences in a genotype network for both protein and RNA molecules. These distances are again smaller for proteins than for RNA. However, in contrast to genotype sets, RNA genotype networks do not traverse genotype space completely.

In sum, the sets of sequence forming any one structure differ between model protein and RNA molecules. Protein genotype sets and networks are

smaller, and extend less far through sequence space than RNA genotype sets, which may traverse genotype space completely.

Shape space covering.

From an evolutionary perspective, genotype networks are important, because they allow genotypic (sequence) change without phenotypic (structure) change. A genotype's neighborhood – all sequences that differ from it in one monomer – may contain different novel phenotypes, depending on the genotype's location on a genotype network (13, 29-30). Thus, genotype networks may facilitate the exploration of novel phenotypes by evolving populations. Larger genotype networks may allow the exploration of more novel phenotypes than small genotype networks (31).

Past computational studies on RNA molecules have uncovered a peculiar feature of RNA secondary structures that has been called *shape space covering* and that has implications for phenotypic evolution (21, 32). For example, a ball of merely r≤15 changed nucleotides around RNA genotypes of length L=100, contains all frequent RNA secondary structures (9, 21), despite the fact that this ball comprises only a vanishing fraction (10^{-37th}) of sequence space. Some work on the HP lattice model indicates that this property may be less pronounced or absent in proteins (4).

We took advantage of the direct comparability of the protein and RNA sequence spaces to characterize how shape space covering may differ between protein and RNA. Specifically, we first asked what fraction of all phenotypes is contained in a ball of a given radius around any one genotype.

In this analysis, we initially focused on genotypes chosen at random from genotype space. Figure 3A shows the total percentage of all phenotypes (vertical axis) that can be encountered in a ball of a given radius (horizontal axis) around a genotype. Observations are averaged over 10^3 randomly chosen genotypes. The figure shows that shape space covering is significantly lower in HP25 proteins than in GC25 RNA molecules. For example a ball with radius R=5 nucleotides contains on average 12.3 percent of RNA phenotypes, whereas it contains on average only 1.1 percent of protein phenotypes (Figure 3A). At a radius of 10 monomer changes this ball would cover 72.8 and 38.9 percent of

phenotypes for RNA and proteins, respectively. Figure S3A shows results of a related analysis based on only those structures that are realized by only one sequence and whose genotype network size is therefore also one. In both, RNA and proteins, the results are almost identical to randomly chosen genotypes.

A next analysis examined the phenotypes accessible within a neighborhood of a given radius around an entire genotype network. We focused on genotype networks in the 0.1 percentile of genotype network size, in order to estimate an upper bound on the percentage of new structures reachable from a genotype network. Because the most populated RNA genotype networks are very large, we sampled 10^3 random sequences from each network, and calculated the fraction of all RNA structures that are contained in neighborhoods of various sizes around these sequences. The largest genotype networks of proteins are smaller, which is why we were able to use all sequences on a genotype network for this analysis. The actual number of sequences accessible from a large RNA genotype network may be even greater than we found, because we were able to study only a sample of sequences from such a network. We note that this renders all differences we discuss below between RNA and protein shape space covering conservative.

The results of this analysis are shown in figure 3B. A comparison with figure 3A shows that a greater percentage of structures can be reached from a large genotype network. For example, whereas only 12.3 percent of RNA structures are reachable through no more than R=5 nucleotide changes from a single randomly chosen genotype, 70.2 percent are reachable through no more than 5 changes from a large genotype network. The corresponding percentages are 1.1 and 14.0 for proteins.

Next, we studied the number of accessible phenotypes from an average sized network. For each RNA and proteins, we sampled 10^3 random genotypes, identified their genotype networks and as before, explored the space covering of every sequence in the network. A greater percentage of structures is accessible at any given radius for RNA sequences than for protein sequences (Figure S3B). For example, 23.2 percent of RNA structures but only 2.5 percent of protein structures are accessible within a radius R=5 of the studied genotype networks.

New structures in genotype neighborhoods.

A genotype neighborhood (P_k), or (k-mutant) neighborhood, is the set of sequences that are no more than k point mutations away from a particular sequence. The novel phenotypes that are the most accessible from any one sequence are those that are a single nucleotide change away from this sequence, that is, they are within the immediate (1-mutant) neighborhood of this sequence. The neighborhoods of different genotypes G_1 and G_2 on the same genotype network can contain different new phenotypes. This is important from an evolutionary perspective, because it means that the existence of genotype networks facilitate phenotypic variability (33). Previous studies have analyzed these differences for RNA molecules as a function of the distance D (in nucleotide changes) between G_1 and G_2 (13, 29). One of these studies showed that the diversity of phenotypes occurring in different neighborhoods increases rapidly as the distance D between genotypes increases (29).

We here wanted to compare this diversity between protein and RNA molecules. To this end, we studied pairs of genotypes G_1 and G_2 on the same genotype network that differed in D nucleotides. We denote as P_1 and P_2 the set of new structures that are found in the neighborhoods of G_1 and G_2, respectively. We were especially interested in the fraction f_D of these structures that occurred in the neighborhood of one but not the other genotype, i.e., we determined f_D=1-$|P_1 \cap P_2|/|P_1|$, where |X| denotes the number of elements in the set X. Note that this analysis of ours is restricted to sequences on the same genotype network. Thus, the maximally possible distance D between the pairs of genotypes we analyze is dictated by the diameter of the genotype network that they are a part of (Figure S2). We studied f_D for genotype networks whose size was in the top 0.1 percentile of all genotype networks. Since large RNA genotype networks may contain thousands of sequences, we only sampled 10^3 genotypes from each genotype network, and calculated f_D for all pairwise combinations of these genotypes. For proteins, we calculated f_D for all protein pairs on a genotype network.

Figure 4 shows the results of this analysis. We note three general features. First, at all distances D between two genotypes, a majority of new structures that occur in one neighborhood do not also occur in the other neighborhood (f_D>0.5).

Second, the fraction f_D of unique structures is statistically indistinguishable between proteins and RNA for D<9, partly because it has a large standard deviation, especially for proteins. Third, for D>9, f_D remains close to one for proteins but decreases for RNA, even though it does stay markedly above f_D=0.5. For example, for HP25 proteins f_9>0.99 more than 99 percent of structures found in neighborhoods of genotypes separated by 9-point mutations are unique to one neighborhood, whereas for GC25 RNA sequences f_9=0.89±.0.09.

In the supplementary material (Figures S5 through S9, Tables S2 through S3), we discuss observations from RNA molecules using the other possible 2-letter RNA alphabet, the AU alphabet. These observations show differences to HP25 proteins similar to those observed for GC25 RNA molecules (Figure S9). The one exception is the last analysis we reported here, where AU25 RNA molecules show much lower neighborhood diversity f_D than GC25 RNA molecules. Neighborhood diversity may thus be highly specific to the RNA alphabet.

A comparison to natural RNA and protein molecules

The data we showed thus far reveal consistent differences between the organization of RNA and protein genotype spaces for our model molecules. Ideally, we would like to compare this data to information from natural RNA and protein molecules, but a thorough comparison is currently not yet possible. First, compared to the size of sequence space there are few natural molecules with known sequence and structure, and these known molecules are not necessarily an unbiased sample from sequence space. Second, several systematic analyses that are possible in the small sequence space we study here are currently impossible for natural molecules. These include the exhaustive analyses of a molecule's neighborhood, or an exploration of phenotypes in a specific region of genotype space. Third, many fewer RNA structures than protein structures are known.

Although these limitations are severe and should be kept in mind, information has recently become available, that allows us to compare at least a few features of natural RNA (6) and protein structures. To this end, the panels of Figure 5 plot the sequence identity between two molecules (horizontal axis)

against the similarity of their tertiary structures (vertical axis). Figure 5A shows this relationship for 2,760 protein pairs, and figure 5B shows it for 1,210 RNA pairs, all of which have known tertiary structures (6). For proteins, plots like this have been pioneered by Chothia and Lesk (34). The dashed vertical lines in both figures indicate sequence identities expected for proteins and RNA with random monomer compositions.

The figure demonstrates that protein sequences at any given sequence identity tend to have more conserved structure than RNA sequences. For example, proteins that share between 40 and 50 percent of their amino acids show 96 percent structural similarity on average, whereas RNA sequences at this divergence show only 84 percent structural similarity on average. Also, the greater the differences between two sequences become, the greater the range of structural similarities that their folds can have (Figure 5A).

A second observation is that for RNA, the structural similarity of the most diverged pairs of molecules at any one sequence identity decreases nearly linearly with sequence identity, which gives the data in Figure 5B its nearly-triangular appearance. This is not true for proteins, where even the most diverged structures are highly similar down to approximately 40 percent sequence identity. For example, for proteins at 50 percent sequence identity, structural similarities fall into a narrow interval ranging from 91 to 100 percent, whereas for RNA molecules at 50 percent sequence identity, structural similarities vary much more broadly, that is, between 57 and 96 percent (Figure 5).

Because natural proteins have a much larger monomer alphabet size of A=20 than natural RNAs with A=4, the question arises whether these differences come from the different alphabet sizes. To address this concern, we have recalculated the data in figure 5A for amino acid alphabets of smaller size (Figure S4), including an alphabet of size four (Figure S4C). This analysis largely preserved the shape of the sequence-structure relationship in figure 5A, and thus confirms that alphabet size does not determine the differences in this relationship for proteins and RNA structures.

The second of the two differences we discussed between RNA and proteins is consistent with our earlier observations on shape space covering.

Specifically, our model molecules showed that regions of a given radius around a sequence contain more structures for RNA than for proteins (Figures 3, S3). This observation is consistent with the triangular shape of the data in Figure 5B, which indicates that RNA molecules at any given sequence divergence adopt more diverse structures than protein molecules at the same sequence divergence.

Discussion

Our observations from tractable RNA and protein genotype spaces confirm two well-known commonalities of the relationship between genotype (sequence) and phenotype (structure) from previous work (9, 14, 27). First, many phenotypes are formed by more than one genotype. The genotypes adopting any one phenotype usually form connected networks of genotypes (Table 1 and Figure 1). Second, some phenotypes are adopted by many more phenotypes than others.

The RNA and protein genotype-phenotype relationships also show major differences, which are the main focus of our work. The first of them is that only a small fraction of protein genotypes -- 0.02 percent for the HP25 model – adopts a unique fold. This is not the case for RNA, where most genotypes – 99 percent in the GC25 data – adopt a unique fold (Table 1).

This observation is consistent with available information from real proteins and RNA molecules. Specifically, few random protein sequences fold into well-ordered structures (24, 28). For example, it has been estimated that 20 percent of random protein sequences with 20 amino acids are soluble (35), and that 5 percent of proteins composed of three different amino acids can fold (28). In contrast, for RNA, a large fraction of random sequences collapse into compact secondary structures (25).

A second major difference is that HP25 proteins form many more structures – even though fewer of their sequences fold – than GC25 RNA molecules. This property is likely to arise from the larger number of possible contacts that each monomer can have in a protein. Specifically, while RNA monomers in a secondary structure can have a maximum of one contact per monomer, protein monomers can have between zero and three contacts, even

for the simple two-dimensional lattice proteins that we consider. These differences also exist for tertiary structures of natural RNA and protein molecules. For example, at radii larger than 3.0 Ångstroms around any one nucleotide and amino acid monomer, the number of other monomers one finds is on average for RNA less than half that observed for proteins (our unpublished observations). These differences are caused by the intrinsic structural properties of monomers and how they interact.

Three more differences between RNA and protein follow from the first two differences: The number of genotypes that form a specific phenotype is smaller for proteins, the number of genotypes in any one genotype network is also smaller for proteins; and the average and maximum distances of genotypes with the same phenotype are smaller for proteins. For example, 32 percent of RNA genotype sets contain genotypes with the maximum distance of 25 nucleotide changes, but none do for proteins (Figure 2B and 2D). Thus, genotype sets and genotype networks are more fragmented for proteins than for RNA.

A last and important final difference regards shape space covering (21). A ball of a given radius around an RNA molecule in sequence space contains a larger percentage of phenotypes than a ball of the same radius around a protein molecule. This is not a self-evident consequence of the first two differences. It indicates that genotype networks are highly interwoven in the case of RNA (9), and less so in the case of proteins (36) (as indicated in Figure 3 and S3).

This last difference has tentative support from our comparative analyses of natural RNA and protein molecules, because RNA molecules below a given sequence divergence D are structurally more diverse than protein molecules whose divergence is below D.

This difference also extends to neighborhoods of entire genotype networks. For example, the total percentage of new structures that occur one mutation away from a genotype network is significantly greater for RNA than for proteins. This observation applies to genotype networks of size one, of average size, and of large size. Overall, this last difference means that RNA genotype space may be more conducive to the exploration of new structure phenotypes.

Any study that uses simplified models of phenotype formation like ours has serious limitations. Perhaps the most important limitation comes from the

need to analyze short sequences with a reduced monomer alphabet in order to study genotype space exhaustively. This limitation can cause “finite size effects” on any observations regarding genotype space organization (5, 37). A candidate example regards the maximal genotype distance for genotypes in the same genotype set. This distance is much smaller than one in our analysis for both natural RNA and proteins. In contrast, evidence from computationally predicted longer RNA structures and from experimental data on protein structures shows that this is not the case for longer molecules with a complete monomer alphabet. For instance, many longer RNA secondary structure phenotypes have genotype networks whose members shall maximal sequence divergence (11, 16). Similarly, proteins with the same fold and a likely common ancestor may show little or no amino acid identity (10, 38-39).

With regard to this limitation, we note that the main purpose of our analysis was to compare protein and RNA genotype spaces. The sequence spaces we have analyzed for RNA and protein have the same size. Thus, although finite size effects certainly exist, both spaces will be affected by them to the same extent, thus making a comparison among these spaces possible.

A second limitation comes from assumptions about how the phenotypes we study are formed. Central to any model of macromolecular phenotype formation is the use of energy functions and of monomer alphabet sizes. In this regard we note that the energy functions of our models reflect well-known biophysical principles. The HP model relies on the role of amino acid hydrophobicity in protein folding (3, 40); and the energy function for RNA secondary structure formation is derived from empirical energy calculations (8). The energy function of the protein model used in this study translates into discrete energy values that are not directly comparable to the context-sensitive energy function of the RNA model. Since the folding thermodynamics of both systems cannot be directly compared, we argue that this difference is part of the intrinsic features of these models. Empirical observations on the foldability of RNA and proteins ultimately support the use of these energy functions (25, 26).

Additionally, in this study we adopted a simple definition of foldability, based only on the degeneracy criterion. According to this definition, a foldable protein only requires a unique minimum of energy in a single conformation.

Other definitions of foldability exist that consider additional criteria, such as the energy difference between the native and the next minimal energy conformation (ie. energy gap). We note that the incorporation of this criterion into our current definition of foldability is only likely to reduce the resulting number of protein structures, and thus increase the differences we observe between proteins and RNA.

With respect to alphabet size, we note that using the same alphabet size is essential if one wants to compare the organization of genotype space for two different classes of molecules, because it ensures that the compared spaces have the same dimension. Functional proteins that contain amino acids drawn from a highly reduced alphabet have been successfully designed (41). Similarly, active RNA ribozymes that use two and three monomer alphabet sizes have been created in the laboratory (42-44).

A third limitation results from a strength of our approach, the ability to study genotype space exhaustively. Because known natural molecules represent neither all possible molecules, nor an unbiased sample of those molecules, it is difficult to compare observations from natural molecules with the simpler models we studied. In addition, the amount of experimental information available for natural RNA structures is still very small, so we cannot be confident that any observed differences to proteins will also hold for larger data sets. This problem will not be resolved until it is possible to characterize RNA and protein tertiary structure phenotypes for many sequences with high accuracy, for example with computationally efficient folding methods or with very high throughput experimental techniques.

Until that time we tentatively conclude based on our limited analysis, that RNA genotype spaces are more conducive to evolutionary searches for novel RNA structure phenotypes by exploring small neighborhoods of genotypes and genotype networks. We are aware that a high diversity of easily accessible structure phenotypes does not imply a high diversity of biochemical functions. For example, it is thought that the larger size of the protein monomer alphabet allows proteins to catalyze more biochemical reactions (45-46). However, where structures and their accessibility matter, RNA may be the more versatile molecule. Candidate examples include many RNA molecules encoded in viral and

other genomes, molecules whose secondary structures have regulatory functions (47-48). It is thus perhaps no coincidence that many such RNA molecules are continually being discovered.

Methods

RNA and protein lattice model

We enumerated all RNA sequences of length 25 composed of either AU or GC nucleotides. We determined the minimum free energy fold for each of the 33,554,432 (2^{25}) possible sequences in each set of sequences using the routine RNAfold from the Vienna RNA package (8) with default parameters. We call a sequence foldable if its minimum free energy structure is unique. We refer to the resulting data sets as the AU25 and GC25 data sets, respectively. Statistics on the fraction of foldable sequences are provided in Tables 1, S1, S2 and S3.

We used the method reported by Irbaeck and Troein (2002) (22) to enumerate the all model protein polymers of length 25 on a two-dimensional lattice. This method encodes the conformational space in a set of allowed moves in space and reduces conformations to '*contact sets*', or conserved combinations of contacts between pairs of hydrophobic amino acids. Sequences are folded consecutively into similar contact sets such that information about previously folded sequences can be used to infer the subsequent ones.

Irbaeck and Troein (2002)'s method is based on the classical HP (hydrophobic – polar) model, where only the contacts between hydrophobic monomers (H) contribute to stability. The total energy of a sequence *S* of length *L*, folded into a conformation *C*, is defined as: $E(S,C) = \sum_{i,j,j>i}^{L} \Delta_{ij} U(s_i,s_j)$, where Δ_{ij} is equal to 1 if and only if monomers at position *i* and *j* contact each other, and are not adjacent on the chain; $\Delta_{ij} = 0$ otherwise. $U(s_i,s_j)$ is the energy function of the HP model, where s_i can take one of two values from the monomer alphabet *A*={H,P}. *U*(H,H) equals -1 and is the only monomer interaction that contributes to the total energy of the confirmation *C*. Foldable sequences are those where only a single confirmation has the minimum energy. We refer to the resulting data set as the HP25 data set. Statistics on the total fraction of foldable sequences are provided in Table 1 and S1.

Sequence and structure data

In November 2010 we obtained 1,883 single-chain proteins from the Protein Data Bank (PDB) (49) solved by X-ray crystallography, with resolutions equal or better than 3.0 A, with no ligands, and with sizes between 100 and 200 amino acids. Structural alignments were produced with the software MAMMOTH (50) from a random sample of 10,000 protein pairs. The method (like the one on which our RNA alignments are based) uses the unit vector alignment strategy (51). From our protein sample we obtained 2,760 highly significant alignments. The p-value of an alignment is calculated assuming that the accuracy of random structural alignments follows a Gumbel distribution (50, 52). Data points shown in Figure 5A are highly significant alingments defined as those with a –ln(p-value) greater than 5.0.

We obtained RNA structure information from the supplementary data of (6). These data are composed of 451 structures that correspond to 101,475 alignments produced with the program SARA (53). Data points in Figure 5B (1,210 alignments) are true positive alignments defined as those with a –ln(p-value) > 4.5 (6).

Acknowledgments

We acknowledge support through Swiss National Science Foundation grants 315200-116814, 315200-119697, and 315230-129708, as well as through the YeastX project of SystemsX.ch, and the University Priority Research Program in Systems Biology at the University of Zurich. EF acknowledges support through UZH Forschungskredit.

Supporting citations

References (54-56) appear in the Supporting Material.

References

1. Wilson, D.S., and J.W. Szostak. 1999. In vitro selection of functional nucleic acids. Annu. Rev. Biochem. 68: 611-647.
2. Maynard Smith, J. 1970. Natural selection and the concept of a protein space. Nature 225:563-564.

3. Lau, K.F., and K.A. Dill. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules. 22:3986–3997.
4. Bornberg-Bauer, E. 1997. How are model protein structures distributed in sequence space? Biophys J. 73:2393-2403.
5. Buchler, N.E.G., and R.A. Goldstein. 1999. Effect of alphabet size and foldability requirements on protein structure designability. Proteins 34:113-124.
6. Capriotti, E., and M. Marti-Renom. 2010. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. BMC Bioinformatics 11:322.
7. Fontana, W., D.A.M. Konings, P.F. Stadler, and P. Schuster. 1993. Statistics of RNA secondary structures. Biopolymers 33:1389-1404.
8. Hofacker, I.L., W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast Folding and Comparison of RNA Secondary Structures. Monatsh. Chem. 125:167-188.
9. Schuster, P., W. Fontana, P. Stadler, and I.L. Hofacker. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc. R. Soc. London B. 255:279-284.
10. Aronson, H.G., E.R. William, and W.A. Hendrickson. 1994. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci. 3:1706-1711.
11. Huynen, M.A. 1996. Exploring phenotype space through neutral evolution. J. Mol. Evol. 43:165-169.
12. Babajide, A., I.L. Hofacker, M.J. Sippl, and P.F. Stadler. 1997. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. Fold Des. 2:261-269.
13. Fontana, W., and P. Schuster. 1998. Continuity in evolution: On the nature of transitions. Science 280:1451-1455.
14. Gruener, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker and P.Schuster. 1996. Analysis of RNA sequence structure maps by exhastive enumeration I. Neutral networks. Monatsh. Chem. 127:355-374.

15. Hardison, R.C. 1996. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc. Natl. Acad. Sci. USA. 93:5675-5679.
16. Saks, M.E., J.R. Sampson, and J. Abelson. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. Science. 279:1665-1670.
17. Dill, K.A., S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. 1995. Principles of protein folding – a perspective from simple exact models. Protein Sci. 4:561-602.
18. Chan, H.S., and E. Bornberg-Bauer. 2002. Perspective on protein evolution from simple exact models. Appl. Bioinformatics. 1:121-144.
19. Schuster, P., and P.F. Stadler. 2004. Discrete models of biopolymers. In Handbook of Computational Chemistry and Biology. M. James, C. Crabbe, and A. Konopka, editors. Marcel Dekker, New York. 187–221.
20. Lipman, D.J., and W.J. Wilbur. 1991. Modelling neutral and selective evolution of protein folding. Proc. R. Soc. London. B. 245:7-11.
21. Gruener, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker and P.Schuster. 1996. Analysis of RNA sequence structure maps by exhastive enumeration II. Structures of neutral networks and shape space covering. Monatsh. Chem. 127:375-389.
22. Irbaeck, A., and C. Troein. 2002. Enumerating Designing Sequences in the HP Model. J. Biol. Phys. 28:1-15.
23. Davidson, A.R., K.J. Lumb, and R.T. Sauer. 1995. Cooperatively folded proteins in random sequence libraries. Nat. Struct. Biol. 2:856-864.
24. Hecht, M.H., A. Das, A. Go, L.H. Bradley, and Y. Wei. 2004. De novo proteins from designed combinatorial libraries. Protein Sci. 13:1711-1723.
25. Schultes, E.A., A. Spasic, U. Mohanty, and D.P. Bartel. 2005. Compact and ordered collapse of randomly generated RNA sequences. Nat. Struct. Biol. 12:1130-1136.
26. Thirumalai, D., and C. Hyeon. 2005. RNA and protein folding: common themes and variations. Biochemistry. 44:4957–4970.
27. Li, H., R. Helling, C.H. Tang, and N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. Science. 273:666-669.

28. Davidson, A.R., and R.T. Sauer. 1994. Folded proteins occur frequently in libraries of random amino acid sequences. Proc. Natl. Acad. Sci. USA. 91:2146-2150.
29. Sumedha, O.C. Martin, and A. Wagner. 2007. New structural variation in evolutionary searches of RNA neutral networks. Biosystems. 90:475-485.
30. Ferrada, E., and A. Wagner. 2010. Evolutionary innovations and the organization of protein functions in genotype space. PLoS ONE. 5:e14172.
31. Wagner, A. 2008. Robustness and evolvability: a paradox resolved. Proc. Biol. Sci. London. B. 275:91-100.
32. Fontana, W. 2002. Modelling evo-devo with RNA. BioEssays. 24:1164–1177.
33. Wagner, A. 2011. The origins of evolutionary innovations. A theory of transformative change in living systems. Oxford University Press.
34. Chothia, C., and A.M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823-826.
35. Prijambada, I.D., T. Yomo, F. Tanaka, T. Kawama, K. Yamamoto, A. Hasegawa, Y. Shima, S. Negoro, and I. Urabe. 1996. Solubility of artificial proteins with random sequences. FEBS Lett. 382:21-25.
36. Bornberg-Bauer, E., and H.S. Chan. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc. Natl. Acad. Sci. USA. 96:10689-10694.
37. Sander, C., and R. Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 9:56-68.
38. Grishin, N.V. 2001. Fold change in evolution of protein structures. J. Struct. Biol. 134:167-185.
39. Nagano, N., C.A. Orengo, and J.M. Thornton. 2002. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J. Mol. Biol. 321:741-765.
40. Kauzmann, W. 1958. Some factors in the interpretation of denaturation. Adv. Protein Chem. 14:1-57.

41. Riddle, D.S., J.V. Santiago, S.T. Bray-Hall, N. Doshi, V.P. Grantcharova, Q. Yi, and D. Baker. 1997. Functional rapidly folding proteins from simplified amino acid sequences. Nat. Struct. Biol. 4:805-809.
42. Reader, J.S., and G.F. Joyce. 2002. A ribozyme composed of only two different nucleotides. Nature 420:841-844.
43. Schlosser, K., and Y. Li. 2009. DNAzyme-mediated catalysis with only guanosine and cytidine nucleotides. Nucleic Acids Res 37(2): 413–420 (2009).
44. Rogers, J., and G.F. Joyce. 1999. A ribozyme that lacks cytidine. Nature. 402:323-325.
45. Qi, D., C. Tann, D. Haring, and M.D. Distefano. 2001. Generation of new enzymes via covalent modification of existing proteins. Chem. Rev. 101:3081–3112.
46. Hendrickson, T.L., V. de Crecy-Lagard, and P. Schimmel. 2004. Incorporation of nonnatural amino acids into proteins. Annu. Rev. Biochem. 73:147-176.
47. Cuceanu, N.M., A. Tuplin, and P. Simmonds. 2001. Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3' end of the hepatitis G virus/GB-virus C genome. J. Gen. Virol. 82:713-722.
48. Thurner, C., C. Witwer, I.L. Hofacker, and P.F. Stadler. 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. 85:1113-1124.
49. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res. 28: 235–242.
50. Ortiz, A.R., C.E.M. Strauss, and O. Olmea. 2002. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci. 11: 2606–2621.
51. Kedem, K., L. Chew, and R. Elber. 1999. Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. Proteins. 37: 554–564.
52. Gumbel, E. 1958. Statistics of extremes. Columbia University Press, New York.

53. Capriotti, E., and M.A. Marti-Renom. 2008. RNA structure alignment by a unit-vector approach. Bioinformatics. 24:112-118.
54. Sharma, P., S. Sharma, A. Mitra, and H. Singh. 2007. Base pairing in RNA structures: A computational analysis of structural aspects and interaction energies. J. Chem. Sci. 119:525-531.
55. Etchebest, C., C. Benros, A. Bornot, A.C. Camproux, and A.G. de Brevern. 2007. A reduced amino acid alphabet for undestanding and designing protein adaptation to mutation. Eur. Biophys. J. 36:1059-1069.
56. Murphy, L.R., A. Wallqvist, and R.M. Levy. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13:149-152.

Tables

Table 1. General statistics of RNA and protein sequence-structure maps.

Model	Uniquely foldable	Number of structures	Foldable fraction
HP25	765,147	107,336	0.023
GC25	33,544,758	31,727	0.999

Table 2. General statistics of RNA and protein genotype networks and genotype sets.

Model	Genotype Sets		Genotype Networks		Networks per set
	Total sets	Sequences per set	Total networks	Sequences per network	
HP25	107,336	7.1 (11.8)	148,254	5.1(9.8)	1.3 (0.7)
GC25	31,727	1,057 (4,827)	2,263,944	14.8 (43.8)	71 (151)

Figures

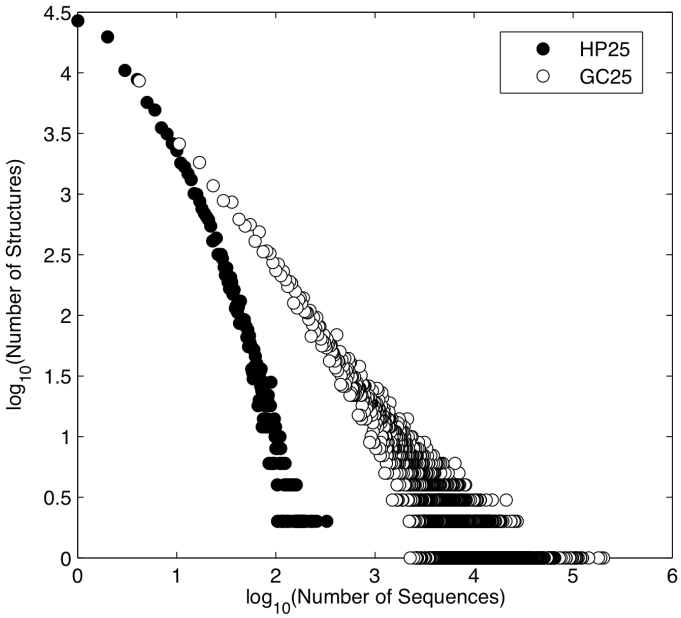


Figure 1. There are many fewer sequences per structure in proteins than in RNA. The figure shows the distribution of the number of structures (vertical axis) that are formed by a given number of sequences (horizontal axis) for the HP25 and GC25 data set. Note the double-logarithmic scale. Data was obtained from exhaustive enumeration of RNA sequences composed of GC nucleotides, and HP protein sequences. Statistics on the number of sequences and structures are presented in Table1.

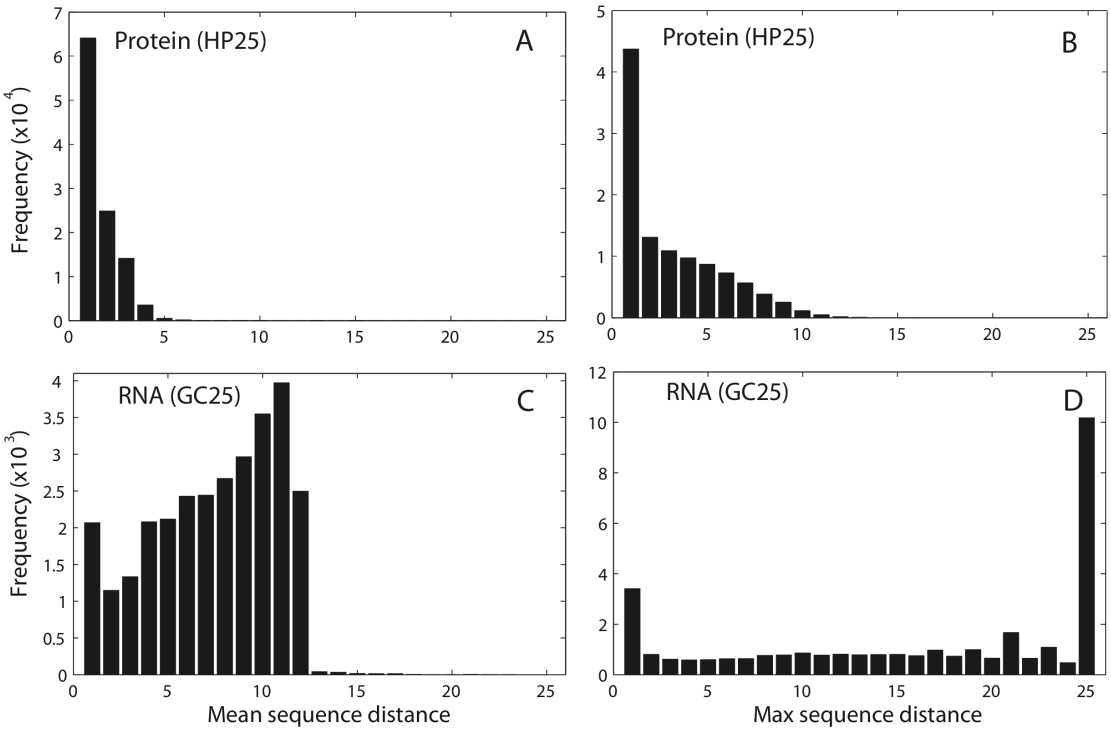


Figure 2. Distribution of the mean and maximum distances of sequences in a genotype set. Plots at the left show the distribution of the mean sequence distances (in number of monomer changes) observed per genotype set in the (A) HP25 and (C) GC25 data. Plots at the right show the distributions of the maximum sequence distance between sequences in the same genotype set, for the (B) HP25 and (D) GC25 data sets.

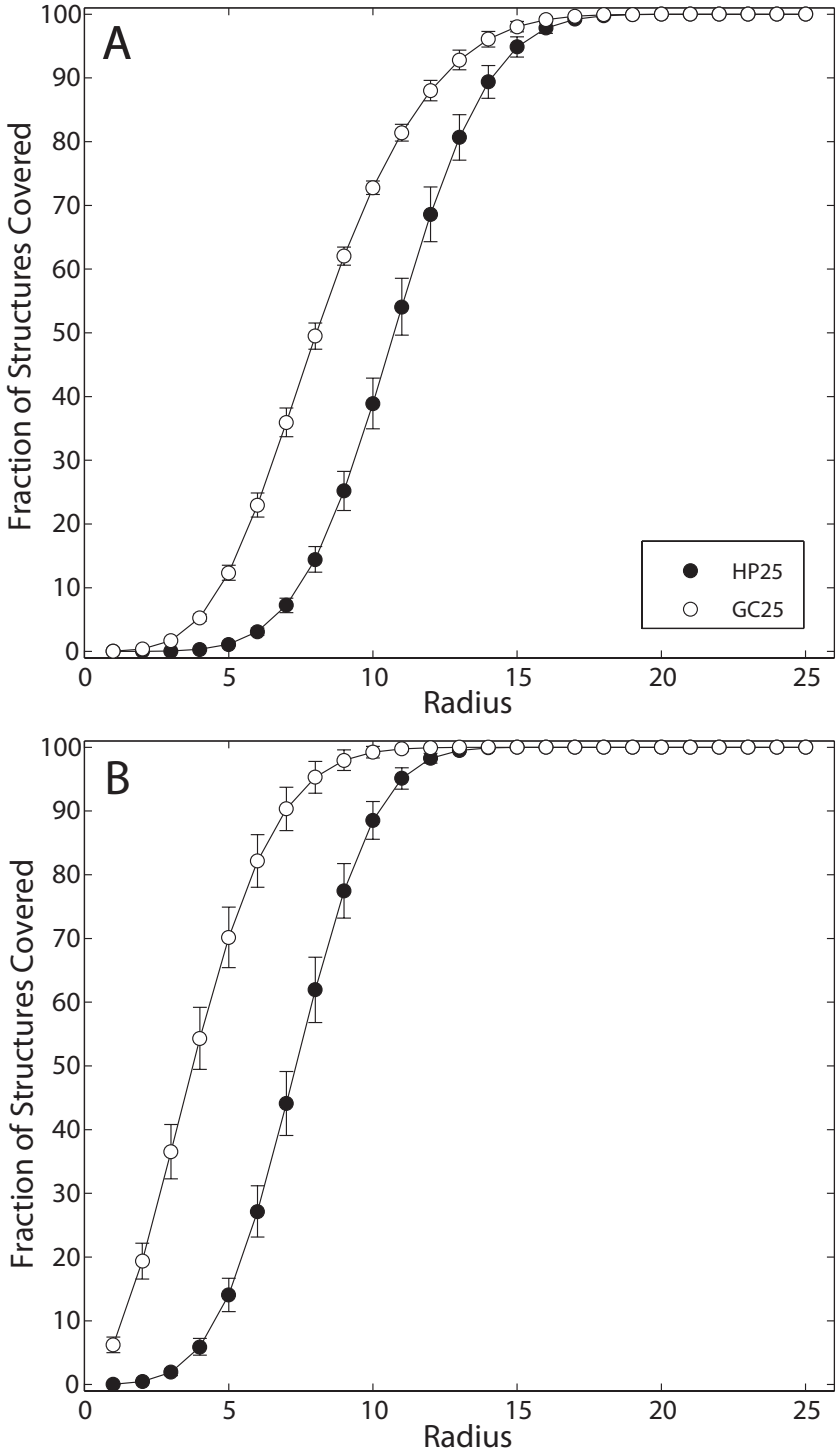


Figure 3. Shape space covering of short RNA and protein sequences with a binary alphabet. A. Shape space covering in neighborhoods of 10^3 sequences sampled at random from genotype space, regardless of the size of the genotype network they belong to. To estimate shape space covering of a particular sequence we determined the percentage of all structures that can be observed within a ball of a given radius (horizontal axis) around the sequence. B. Shape

space covering of the most populated genotype networks. We estimate the shape space covering of an entire network by counting the number of different phenotypes contained within a neighborhood of a given radius around every sequence in the network. The data shown are based on all genotype networks in the top 0.1 percentile of genotype network size. This percentile corresponds to 2,260 and 148 RNA and protein genotype networks, respectively. Error bars correspond to one standard deviation.

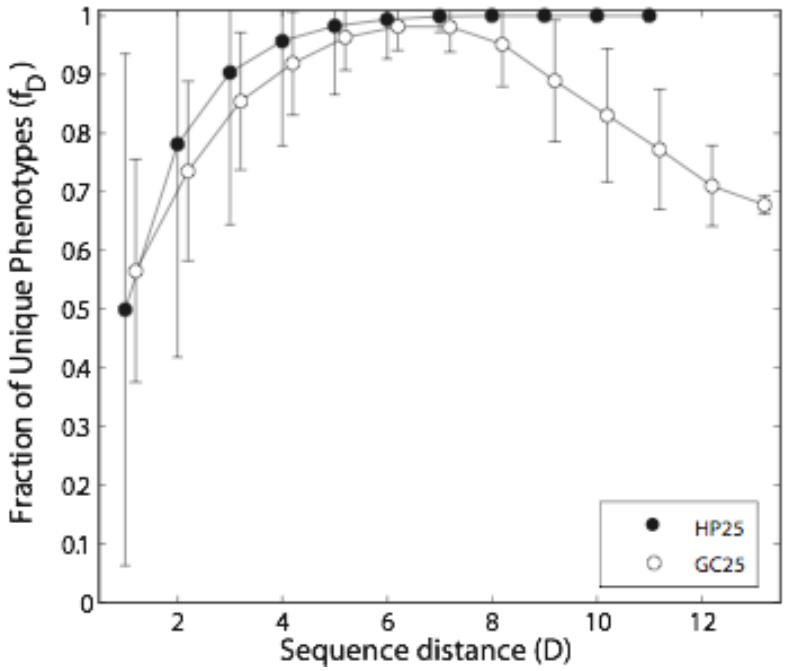


Figure 4. Unique novel structures in the neighborhood of different genotypes on the same genotype network. The horizontal axis shows the genotype distance *D* between two genotypes on the same genotype network. The vertical axis shows the fraction of new phenotypes (f_D) that is unique to one neighborhood, in the sense that it occurs in the neighborhood of one of these genotypes but not the other. Data is based on genotype networks in the top 0.1 percentile of genotype network size. See main text for details. Error bars correspond to one standard deviation.

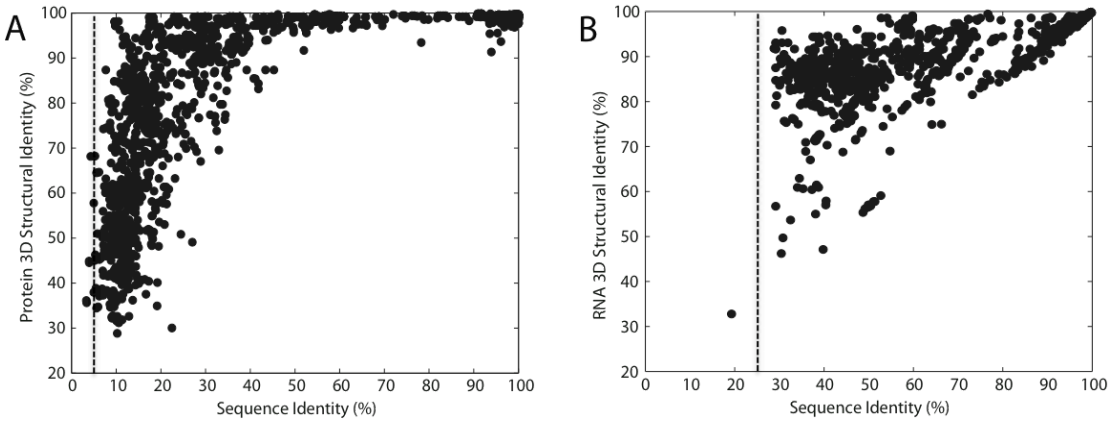


Figure 5. A comparison of sequence-structure relationships for natural proteins and RNA molecules. A. Sequence identity versus tertiary structure identity for proteins. The figure shows sequence identity calculated over the structurally aligned residues (horizontal axis) versus structural identity (vertical axis). The figure is based on pairwise comparisons of 1,883 single-chain proteins from PDB (49) that were solved by X-ray crystallography and that fulfilled the following criteria: The structure's resolution is at least 3.0 A, the protein has no bound ligands, and a size that lies between 100 and 200 amino acids. Structural alignments were produced with the software MAMMOTH (50) from a random sample of 2,760 protein pairs (see Methods). Data points shown in Figure 5A were filtered at a logarithmically (base e) transformed p-value exceeding 5.0. B. Sequence identity versus tertiary structure identity for RNA. The figure shows sequence identity over all structurally aligned residues (horizontal axis) versus percentage of structural identity (vertical axis). The data is based on 1,210 alignments (158 structures) extracted from a larger data set of 451 structures with 101,475 alignments produced with the program SARA (53).

2.1 Supplementary Material

In this section we explore the effects of two different binary alphabets on the RNA genotype-phenotype map. As shown in the main text, the HP25 protein and GC25 RNA models show extensive difference in the number of phenotypes and the fraction of foldable sequences (Table 1). Here, we report and compare analogous statistics for RNA sequences with L=25 nucleotides drawn from either the GC or the AU alphabet. Table S2 shows that the AU alphabet produces fewer uniquely foldable sequences, and that its repertoire of structures is smaller than for the GC alphabet. Additionally, the fraction of foldable sequences is half that observed for the GC alphabet (Table S2).

The distribution of the number of AU sequences per structure

The distribution of the number of sequences that fold into any one structure is very similar for both the AU25 and the GC25 data sets (Figure S5). The number of sequences per structure shows a non-uniform distribution, with a marked predominance of structures adopted by few sequences (Figure S5). Table S3 shows pertinent summary statistics from exhaustive enumeration. The GC25 data set contains 20 times more networks, but they are on average 10 times smaller than in the AU25 data set.

The RNA alphabet affects the total number of RNA structures, which may be explained by differences in the energetic contribution of base pair interactions. Specifically the approximate 3.6-fold increase in the free energy of AU interactions compared to GC interactions (1) translates into a 50-fold decrease in the number of conformations we estimate for the AU25 data set (Table S3).

Figure S6 shows the distributions of the average sequence distance between sequences of the same genotype set. The AU25 and GC25 datasets show similar mean sequence distances (8.7±3.1; 7.4±3.3, respectively). Mean sequence distances do either not exceed 12 nucleotide changes (AU25) or they rarely do (for 0.4 percent of structures in the GC25 genotype sets) (Figure S6). The distributions of maximum distances between sequences of the AU25 and GC25 models are shown in Figure S7. 65 percent and 44 percent of genotype sets show

maximum distances larger than 20-point mutations for the AU25 and GC25 models, respectively. Moreover, as for the GC25 data set, many genotype sets of the AU25 model span genotype space completely. Indeed, 44 percent of AU25 genotype sets (and 32 percent of GC25 genotype sets) show the maximum distances of 25. As discussed in the main text, any one genotype set may be composed of more than one connected component or genotype network. The distribution of the number of genotype networks per genotype set differs between the AU25 and GC25 models (Table S3). Figure S7 shows the distributions of maximal and mean distances between pairs of sequences that belong to the same genotype network.

Shape space covering

As shown in the main text, balls of a given radius centered on a sequence contain a greater percentage of structures for RNA than for protein. This extent of shape space covering is even greater for AU sequences then for GC sequences, as Figure S8A shows. For example, a ball with a radius of 4-point mutations around an AU sequence covers on average 25 percent of all RNA structures, while such a ball covers only 5 percent of RNA structures for GC25 sequences. A ball with a radius of 7-point mutations, roughly the average distance observed between randomly generated sequences, would contain 69 percent of all structures for AU sequences, but only 36 percent for GC sequences. (These values are based on sequences that belong to the smallest genotype networks of size 1). Figure S8B shows analogous statistics, but for genotype networks that are in the top 0.1 percentile of genotype network size.

Phenotypic neighborhood diversity in AU25 genotype networks

Figure S9 shows the fraction of unique phenotypes f_D in 1-mutant neighborhoods around pairs of sequences at genotype distance D (see main text for details). The figure shows that f_D is lower for AU sequences than for GC sequences for all but the largest sequence distances we considered. For example, while the GC25 model attains over 95 percent of unique new phenotypes (f_D=0.95) for neighborhoods of genotypes that are only D=5 point mutations apart, the AU25 model reaches no more than 60 percent of unique new

phenotypes at any genotype distance *D* (Figure S9). This implies that neighborhood phenotypic diversity is highly sensitive to the nucleotide alphabet in our model sequences.

Supporting References

1. Sharma, P., S. Sharma, A. Mitra, and H. Singh. 2007. Base pairing in RNA structures: A computational analysis of structural aspects and interaction energies. J. Chem. Sci. 119:525-531.
2. Riddle, D.S., J.V. Santiago, S.T. Bray-Hall, N. Doshi, V.P. Grantcharova, Q. Yi, and D. Baker. 1997. Functional rapidly folding proteins from simplified amino acid sequences. Nat. Struct. Biol. 4:805-809.
3. Murphy, L.R., A. Wallqvist, and R.M. Levy. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13:149-152.
4. Etchebest, C., C. Benros, A. Bornot, A.C. Camproux, and A.G. de Brevern. 2007. A reduced amino acid alphabet for undestanding and designing protein adaptation to mutation. Eur. Biophys. J. 36:1059-1069.

Table S1. Statistics on genotype set.

Model	Total Sets	Networks of size one	Sets with more than one neutral network
HP25	107336	87567	30536
AU25	648	647	590
GC25	31727	29382	27080

Table S2. General statistics of RNA and protein sequence-structure maps.

Model	Uniquely foldable sequences	Number of structures	Foldable fraction
AU25	13,643,201	648	0.407
GC25	33,544,758	31,727	0.999

Table S3. General statistics of RNA and protein genotype sets.

Model	Neutral Sets		Neutral Nets		Networks per set
	Total sets	Sequences per set	Total networks	Sequences per network	
AU25	648	21,054 (39,047)	93,992	145 (342)	145 (214)
GC25	31,727	1,057 (4,827)	2,263,944	14.8 (43.8)	71 (151)

Table S4. Reduced amino acid alphabets used in Figure S4.

Figure	Alphabet size	Amino acid groups	Reference
Figure S4A	20	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y	
Figure S4B	5	AHT, ED, GP, IVFYWLMC, KNQRS	Ridley et al 1997
Figure S4C	4	AGPST, CILMV, DEHKNQR, FYW	Murphy et al 2000
Figure S4D	5	G, IVFYW, ALMEQRK, P, NDHSTC	Ethebest et al 2007
Figure S4E	8	G, IV, FYW, ALM, EQRK, P, ND, HSTC	Ethebest et al 2007
Figure S4F	9	G, IV, FYW, ALM, EQRK, P, ND, HS, TC	Ethebest et al 2007
Figure S4G	11	G, IV, FYW, A, LM, EQRK, P, ND, HS, T, C	Ethebest et al 2007
Figure S4H	13	G, IV, FYW, A, L, M, E, QRK, P, ND, HS, T, C	Ethebest et al 2007

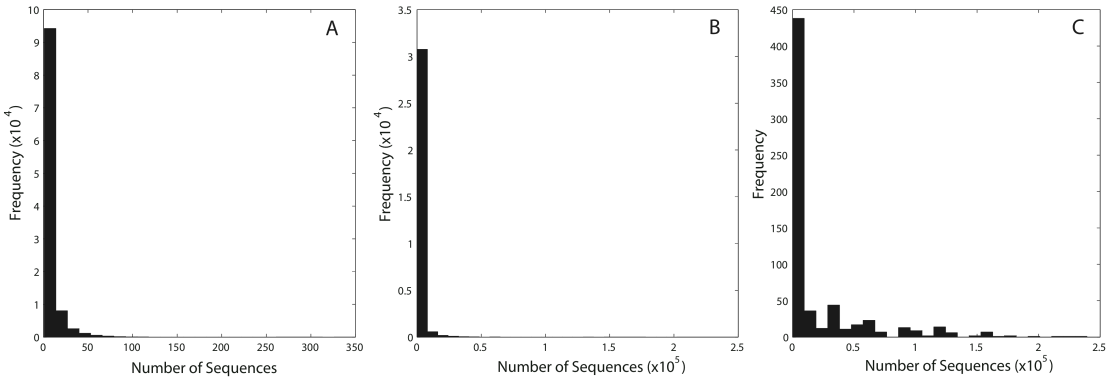


Figure S1. Histograms of the number of sequences per genotype set of the HP25, GC25, and AU25 data sets. For each data set, exhaustive enumeration is performed and the number of sequences folding into each conformation is counted.

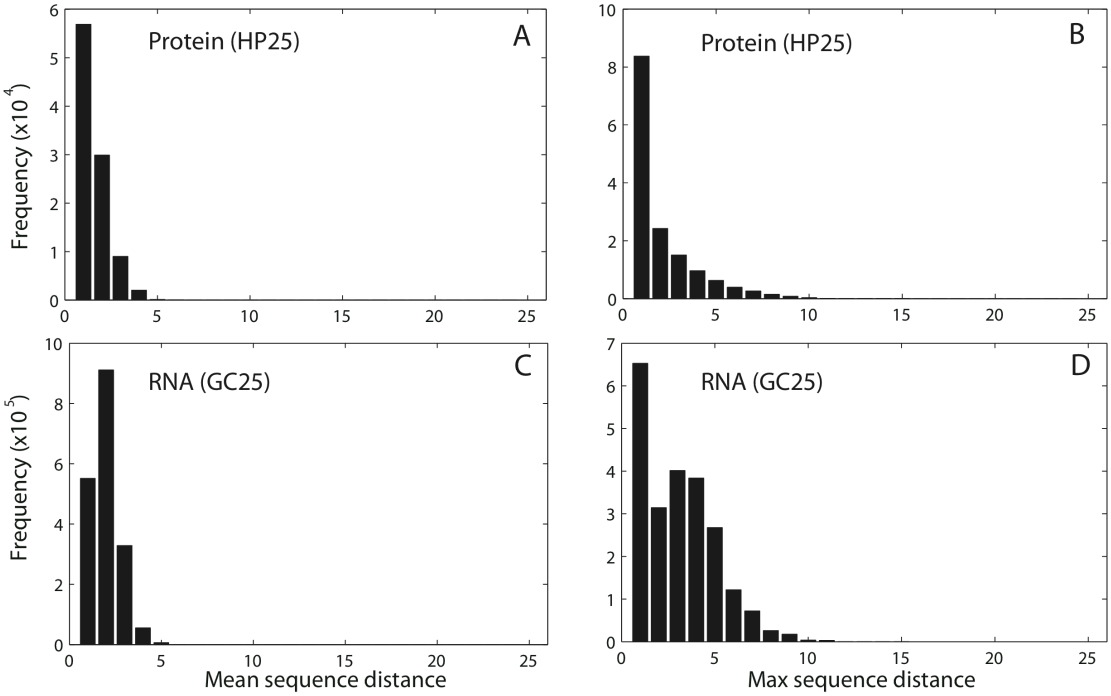


Figure S2. Distribution of the mean and maximum sequence distances per genotype network. Plots at the left show distributions of mean distances among sequences in the same genotype network for (A) HP25 proteins and (C) GC25 RNA. Plots at the right show distributions of the maximum sequence distance between sequence pairs in the same genotype network, for (B) HP25 proteins and (D) GC25 RNA.

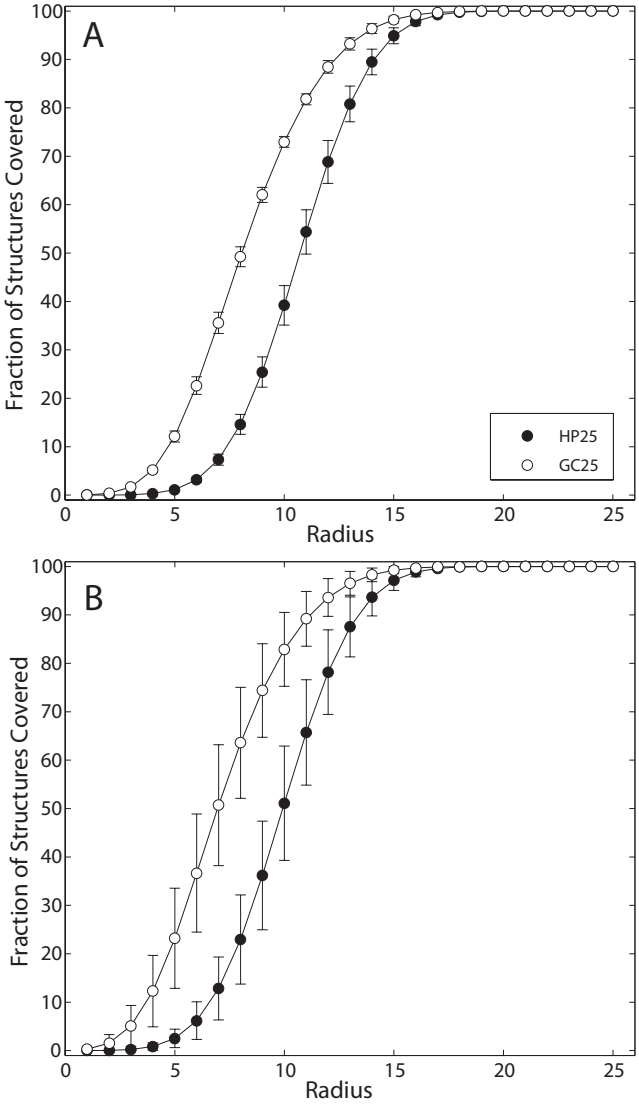


Figure S3. Shape space covering of short RNA and protein sequences with a binary alphabet. A. Shape space covering of 10^3 randomly sampled genotype "networks" of size 1. To estimate the shape space covering of a particular sequence we determined the percentage of all structures observed within a ball of a given radius (horizontal axis) around the sequence. B. Shape space covering of typical genotype networks. We calculated the shape space covering of an entire genotype network by determining the percentage of all phenotypes contained within a neighborhood of a given radius around every sequence in the network. Specifically, we sampled 10^3 genotypes at random, determined this percentage for the genotype network that each genotype is a part of, and show averages of this percentage over the 10^3 genotypes (vertical axis). Error bars correspond to one standard deviation.

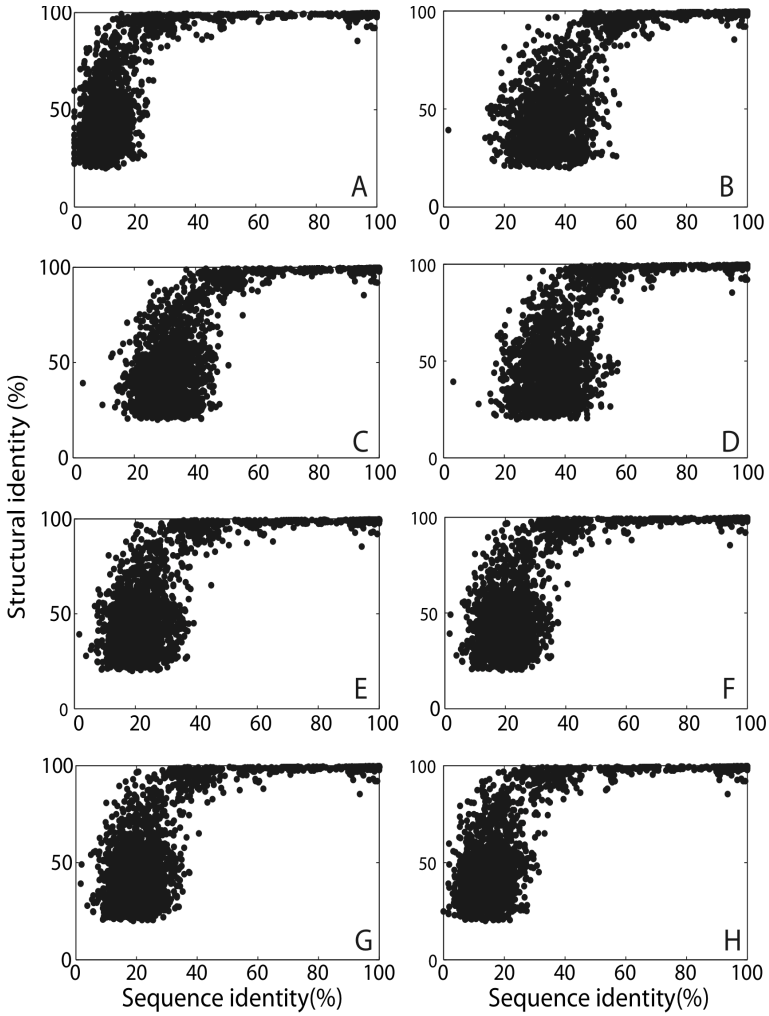


Figure S4. Reduced amino acid alphabet size does not dramatically change the relationship between structural similarity and sequence similarity for natural proteins. We use the same protein data set as described in the caption to Figure 5A. We determined structural alignments with the software MAMMOTH (50). We only analyzed structure alignments further that were at least 50 amino acids long. We used each structure alignment to calculate sequence identity by replacing each amino acid with an amino acid taken from a reduced amino acid alphabet. We note that the algorithm implemented in MAMMOTH does not use sequence information in the structural alignment, thus rendering our procedure of obtaining reduced amino acid alphabets unproblematic. We used the following amino acid alphabets: A) the standard alphabet (A=20); B) an alphabet proposed by Ridley et al (1997) (A=5); C) an alphabet A=4, proposed by Murphy et al (2000). Panels D to H are based on alphabets proposed by Etchebest et al (2007). D) A=5 ; E) A=8; F) A=9; G) A=11 and H) A=13. Alphabets and references are detailed in Table S4.

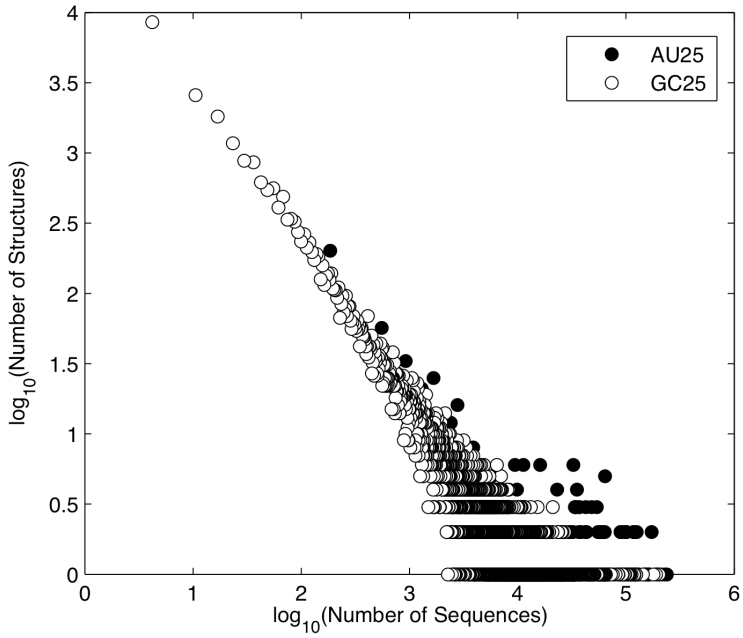


Figure S5. The distribution of sequences per structure in the AU25 and GC25 RNA models. The figure shows the distribution of the number of structures (vertical axis) that are formed by a given number of sequences (horizontal axis) for the AU25 and GC25 data set. Note the double-logarithmic scale. Data was obtained from exhaustive enumeration of RNA sequences containing only AU or GC nucleotides. Statistics on the number of sequences and structures are presented in Table S2.

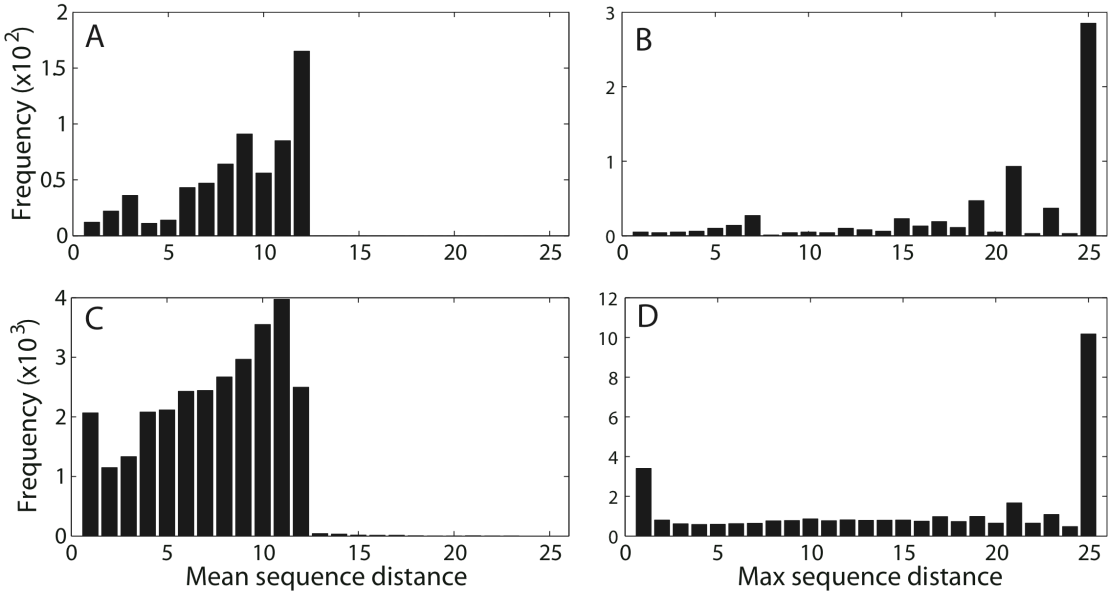


Figure S6. Distribution of the mean and maximum distances of sequences per genotype set. Plots at the left show the distribution of the mean sequence distances (in number of monomer changes) observed per genotype set in the (A) AU25 and (C) GC25 data. Plots at the right show the distributions of the maximum sequence distance between sequences in the same genotype set, for the (B) AU25 and (D) GC25 data sets.

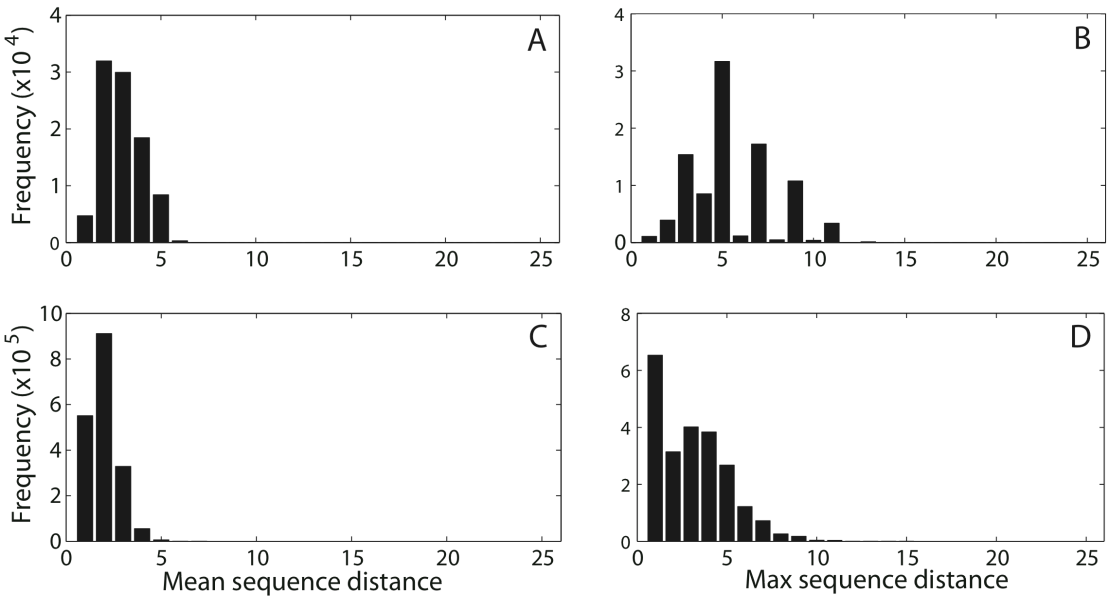


Figure S7. Distribution of mean and maximum sequence distances per genotype network. Plots at the left show distributions of mean distances among sequences in the same genotype network in the (A) AU25 and the (C) GC25 RNA data set. Plots at the right show distributions of the maximum sequence distance between sequence pairs in the same genotype network in the (B) AU25 and the (D) GC25 RNA data set.

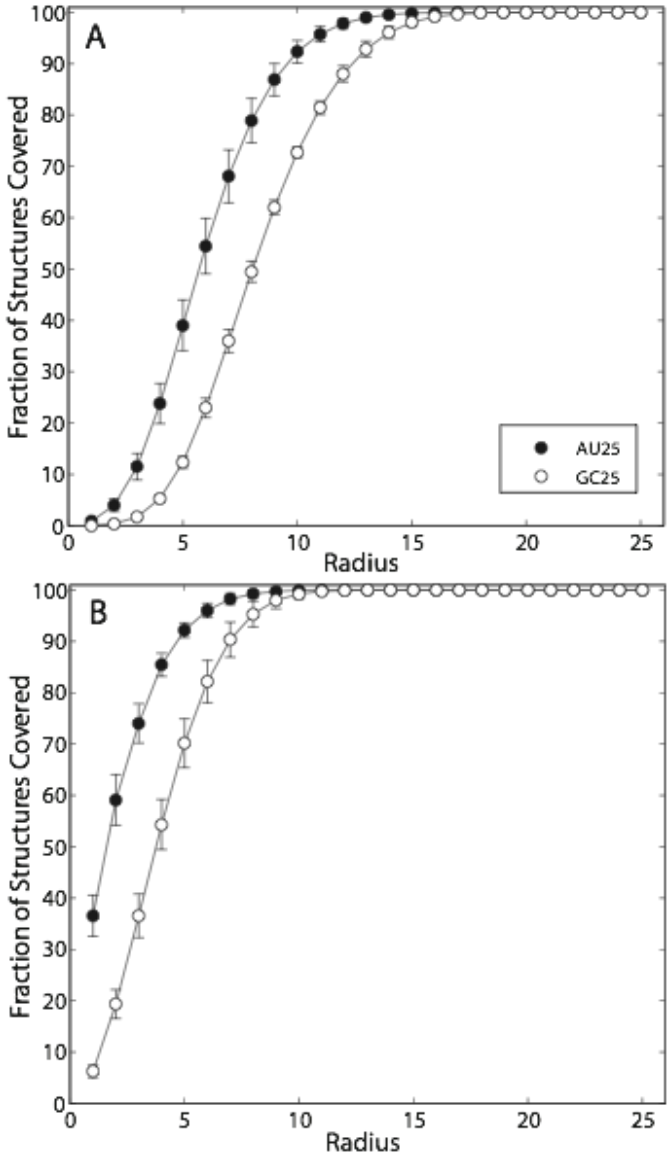


Figure S8. Shape space covering of short RNA sequences with AU and GC nucleotides. A. Average shape space covering of 10^3 randomly sampled sequences, regardless of the size of the genotype network they belong to. To estimate the shape space covering of a particular sequence, we determined the percentage of structures observed within a ball of a given radius (horizontal axis) around the sequence (see main text). B. Shape space covering of the most populated genotype networks. We calculated the shape space covering of an entire genotype network by determining the percentage of phenotypes contained within a neighborhood of a given radius around every sequence in the network. Panel B shows this quantity for genotype networks in the 0.1 percentile of genotype network size, for both the AU25 and GC25 data. Error bars correspond to one standard deviation.

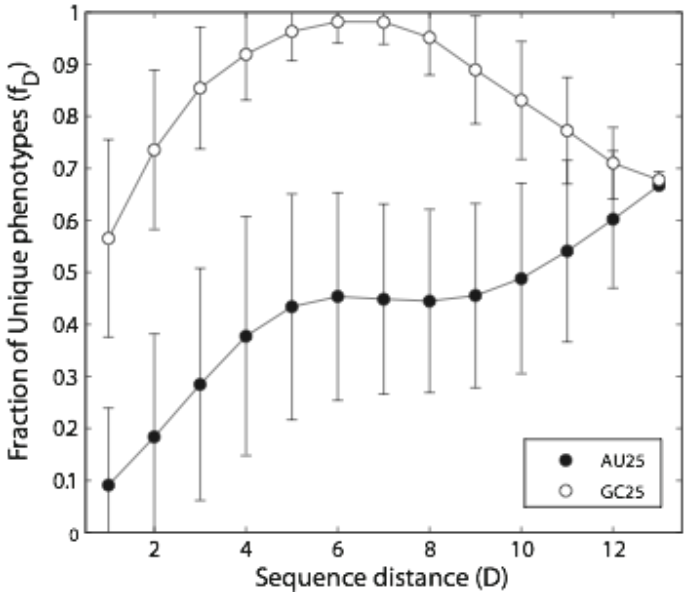


Figure S9. Unique novel structures in the neighborhood of different genotypes on the same genotype network. The horizontal axis shows the genotype distance D between two genotypes on the same genotype network. The vertical axis shows the fraction of new phenotypes (f_D) unique to the neighborhood of one of these genotypes. Data is based on genotype networks in the top 0.1 percentile of genotype network size, for both the AU25 and GC25 RNA data sets. For each genotype network, we performed all-against-all comparisons among the sequences in the sample, and f_D was calculated separately for all genotype pairs at a given distance, as specified in the main text. Error bars correspond to one standard deviation.

3. Protein robustness promotes evolutionary innovations on large evolutionary time-scales

Ferrada E and Wagner A. Protein robustness promote evolutionary innovation on large evolutionary time-scales. *Proc Biol Sci Lond B* 275:1595-1602 (2008).

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 1595–1602
doi:10.1098/rspb.2007.1617
Published online 22 April 2008

Protein robustness promotes evolutionary innovations on large evolutionary time-scales

Evandro Ferrada[1,3,*] and Andreas Wagner[1,2,3]

[1]*Department of Biochemistry, University of Zurich, Building Y27, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[2]*The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[3]*The Swiss Institute of Bioinformatics*

Recent laboratory experiments suggest that a molecule's ability to evolve neutrally is important for its ability to generate evolutionary innovations. In contrast to laboratory experiments, life unfolds on time-scales of billions of years. Here, we ask whether a molecule's ability to evolve neutrally—a measure of its robustness—facilitates evolutionary innovation also on these large time-scales. To this end, we use protein designability, the number of sequences that can adopt a given protein structure, as an estimate of the structure's ability to evolve neutrally. Based on two complementary measures of functional diversity—catalytic diversity and molecular functional diversity in gene ontology—we show that more robust proteins have a greater capacity to produce functional innovations. Significant associations among structural designability, folding rate and intrinsic disorder also exist, underlining the complex relationship of the structural factors that affect protein evolution.

Keywords: robustness; evolutionary innovations; protein designability; functional diversity

1. INTRODUCTION

What makes a biological system able to produce evolutionary innovations (Müller & Wagner 1991), new adaptations that may aid in survival and reproduction? Do some systems have a greater ability to innovate than others? A rigorous answer to these questions requires a systematic comparison of many different systems and the innovations they have produced. Whole organisms are not readily amenable to such systematic comparison. By contrast, molecular innovations can be more easily studied. This is because we know millions of protein sequences, as well as thousands of structures, and their associated functions. For this reason, here we address the opening questions with protein molecules and their functional diversity, which is a record of past evolutionary innovations.

Recent experimental work suggests that a molecule's ability to evolve neutrally is important for its ability to evolve new functions. Such neutral evolution leaves a primary function of the molecule unchanged, while paving the way for new functions to emerge. Cases in point are the enzymes serum paraoxonase and cytochrome P450. These enzymes have a primary catalytic function, but they can also metabolize other secondary substrates at greatly reduced rates (Amitai *et al.* 2007; Bloom *et al.* 2007). Laboratory evolution experiments show that neutral mutations that do not change the primary function of these enzymes can cause substantial fluctuations in their secondary activities. Natural selection can then rapidly increase these 'promiscuous' activities (Aharoni *et al.* 2005). A different kind of experiment with two catalytic RNA molecules makes a similar point. In this experiment, Schultes & Bartel (2000) mutagenized two ribozymes unrelated in sequence, structure and catalytic activity. These authors created a path of single mutations through sequence space that connected the two ribozymes. After most of the steps in this path, the catalytic activity of the mutated molecules did not change much, except for a small transition region approximately halfway between the two starting molecules. In this region, the activity of one molecule switched to the activity of the other molecule. Here again, neutral mutations paved the way for a molecule with a new function. In both cases, the ability to evolve neutrally facilitated a molecule's ability to acquire functional innovations.

If these observations hold more generally, the following prediction arises for two different molecules A and B: if A can undergo more neutral mutations than B—it has greater mutational robustness than B—then A should also show a greater propensity to evolve new functions. This prediction has been confirmed for cytochrome P450 in another recent experiment (Bloom *et al.* 2006*b*), which showed that thermostable or mutationally robust variants of this enzyme more readily evolve new catalytic activities. A theoretical work on RNA structures provides a larger context and intuitive explanation for this observation (Wagner 2007). Populations of mutationally robust structures can explore a set of all possible genotypes rapidly through neutral mutations. They are thus genotypically diverse and can produce large amounts of structural variation by single point mutations. This increased access to structural diversity promotes evolutionary innovations, even though only a small fraction of structural variants may lead to new functions.

Laboratory experiments can explore evolutionary innovations on laboratory time-scales. However, life unfolded on time-scales of billions of years. Does the connection between robustness and evolutionary innovation hold on these vastly larger time-scales? This is the question we address here. To do so, we need to analyse a protein

*Author and address for correspondence: Department of Biochemistry, University of Zurich, Building Y27, Winterthurerstrasse 190, 8057 Zurich, Switzerland (e.ferrada@bioc.uzh.ch).

Received 12 March 2008
Accepted 2 April 2008

structure's ability to evolve neutrally—its mutational robustness—for many different structures. This ability is directly related to the number of sequences in a genotype space that can fold into a given structure, also known as the designability of the structure. The concept of designability was first coined by Li *et al.* (1996). Using a simple lattice model, these authors showed that the number of sequences that can adopt a given structure is related to the structure's regularity and to its robustness to mutations. Further studies have shown that designability is also related to evolutionary rate (Bloom *et al.* 2006*a*). The sequences folding into a structure are typically connected in large neutral networks (Babajide *et al.* 1997; Bastolla *et al.* 2003).

Here we show that more robust proteins show greater propensity to evolve new functions on vast evolutionary time-scales. To this end, we use quantitative estimates of protein designability that can be determined from a protein's contact density matrix (England & Shakhnovich 2003), or from the diversity of sequences adopting a protein structure (Shakhnovich *et al.* 2005). As a record of past evolutionary innovations, we use the functional diversity of protein domains, as encapsulated in their diversity of enzymatic functions (Pegg *et al.* 2006) and in their gene ontology annotations (Ashburner *et al.* 2000) of molecular functions.

2. MATERIAL AND METHODS

Our main source of data is the class, architecture, topology and homologous superfamily (CATH) protein structure classification database v. 3.1.0 (Greene *et al.* 2007). Here we focus on the 1924 representative protein domains in CATH, which exceed a minimal length of 50 residues. The number of different functions known for a domain depends on the time since a domain originated in evolution: for two domains—one young and another old—with equal designability (robustness), the young domain had less time to accumulate sequence and functional diversity. We exclude this confounding factor by focusing some of our analyses on a subset of ancient domains that are present in all sequenced bacterial, archaeal and eukaryotic genomes (Ranea *et al.* 2006), and that were thus present in the last universal common ancestor of extant life. Since this dataset was derived from a previous CATH release, we filter these domains to obtain 112 ancient domains that occur in the current release.

(a) *Measures of designability*

In our analysis, we use two complementary estimates of a protein fold's designability. We refer to these estimates as structural designability (D_S) and diversity designability (D_D). Structural designability was introduced by England & Shakhnovich (England & Shakhnovich 2003; Shakhnovich *et al.* 2005). These authors showed that the number of sequences that can adopt a given structure is approximated by the length-normalized maximum eigenvalue of the contact density matrix at a defined distance cut-off, based on a coarse-grained structural description (using only C_α and C_β atoms). The contact density matrix $A=(a_{ij})$ is a binary (0-1) matrix, where $a_{ij}=1$ if two residues i and j that are not neighbours ($|i-j|>1$) are in contact. For our purpose, we consider two non-neighbouring residues in contact, if any of their C_α and C_β atoms occur within a 6.0 Å radius of each other. An alternative measure of structural designability is the average number of atomic contacts per residue (England & Shakhnovich 2003; Bloom *et al.* 2006*a*). However, this measure is so closely correlated with D_S (Spearman's $r=0.989$; $p<10^{-100}$) that it yields virtually identical results. We thus focus exclusively on the length-normalized structural designability, D_S.

We obtain our second estimate of designability (D_D) from diversity data of protein sequences, in an approach similar to that of Shakhnovich *et al.* (2005). Specifically, we analyse sequences in the non-redundant dataset NRDB90 (Holm & Sander 1998). We examine each sequence in this set and assign it to an ancient representative CATH domain, if the sequence has 25% or more identity to the CATH representative, as suggested by the analysis of Chothia & Lesk (1986). We use BLAST (Altschul *et al.* 1997) to determine the extent of sequence identity. Since the number of similar sequences observed per representative domain is dependent on its length, we also normalized D_D by the sequence length.

Because designability may be related to the complexity and amount of disorder of a protein fold, we also explored their relationship with functional diversity. As a measure of fold complexity, we used the absolute contact order (ACO) as introduced by Plaxco *et al.* (1998). ACO is the average distance on the amino acid sequence of two residues that contact each other in the structure. Proteins with high ACO fold slowly. We calculate ACO as in Ivankov *et al.* (2003), where we consider two residues to be in contact if any of their C_α or C_β are inside a sphere of 6.0 Å.

To explore intrinsic disorder (ID) in the sequence domain dataset described above, we use the tool IUPred (Dosztányi *et al.* 2005*a*,*b*). Briefly, IUPred estimates for each residue in a sequence an index that indicates the amount of disorder this residue is subject to. We calculate the disorder average for each sequence in the NRDB90 dataset and assign this value to a CATH representative domain if the BLAST comparison shows a per cent identity of the sequence that is greater than 25%. Finally, we simply calculate the average over the whole set of disorder scores assigned to a representative domain.

(b) *Functional annotation*

We estimate the capacity to evolve functional innovations using information from two sources. The first is the structure–function linkage database (SFLD) that associates sequence, structure and functional annotation for a diverse spectrum of enzyme superfamilies. This functional annotation is based on structural similarities of enzyme active sites (Pegg *et al.* 2006). In September 2007, the SFLD contained 6280 protein sequences grouped in 138 families and six superfamilies. We determined the diversity of functions on the family level for all sequences that shared more than 25% identity with any of the CATH representative domains.

We express functional diversity of a domain in two ways. The first (FE_1) is simply the number of different SFLD families assigned per domain and normalized by the domain length. The length-normalization is needed to correct for the fact that the longer the sequence, the higher the chance to find a second sequence that shares 25% of identity. The second (FE_2) is a measure akin to an entropy that takes into account the frequency of different enzymatic functions observed per domain. If a set of sequences associated with a domain has k different associated enzymatic functions (some of which may occur multiple times), and if p_i is the frequency with which each function i occurs in the set of sequences, then $FE_2 = -\sum_{i=1}^{k} p_i \log p_i$.

Table 1. Spearman's rank correlation coefficients. (D_S, structural designability; D_D, diversity designability; FE_1, enzymatic functional diversity; FG_1, diversity of molecular functions (based on gene ontology); FE_2, entropic measure of enzymatic functional diversity; FG_2, entropic measure of molecular functional diversity (GO); ACO, absolute contact order; ID, intrinsic disorder. The upper right triangle shows Spearman's rank correlation coefficients (r). The lower left triangle shows the corresponding p-values. Diversity designability as well as functional diversity measures are reported for the set of highly conserved evolutionary domains.)

	D_S	D_D	FE_1	FG_1	FE_2	FG_2	ACO	ID
D_S	—	0.882	0.702	0.938	0.877	0.973	−0.698	0.923
D_D	7.25×10^{-53}	—	0.801	0.961	0.877	0.868	−0.662	0.897
FE_1	1.09×10^{-30}	2.31×10^{-40}	—	0.818	0.872	0.700	−0.625	0.705
FG_1	2.99×10^{-68}	2.50×10^{-79}	1.69×10^{-42}	—	0.916	0.938	−0.765	0.931
FE_2	7.13×10^{-52}	7.13×10^{-52}	6.40×10^{-51}	5.42×10^{-61}	—	0.886	−0.604	0.889
FG_2	4.08×10^{-88}	3.48×10^{-50}	1.58×10^{-30}	2.99×10^{-68}	1.09×10^{-53}	—	−0.638	0.952
ACO	9.11×10^{-91}	1.14×10^{-27}	3.56×10^{-25}	2.22×10^{-36}	7.25×10^{-24}	5.07×10^{-26}	—	−0.607
ID	$<10^{-100}$	4.07×10^{-56}	6.23×10^{-31}	1.08×10^{-65}	2.50×10^{-54}	2.29×10^{-74}	$<10^{-100}$	—

The second source of functional information used in this study is the GOA database that maps UniProt (The UniProt 2007) entries to gene ontology (GO) terms (Camon *et al.* 2004). We obtained the GOA database from the EMBL-EBI FTP site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT), and filtered the complete database to obtain only those UniProt entries that were annotated with molecular functions. We then created a non-redundant database of sequences using the NRDB90 tool (Holm & Sander 1998). Subsequently, we examined each sequence in this database and mapped the associated GO terms to a CATH representative domain, if the sequence shared more than 25% identity with the CATH domain. Analogous to enzymatic diversity, we express molecular functional diversity in two ways. The first (FG_1) is simply the number of different GO molecular functions per domain, normalized by the domain length. The second (FG_2) is the entropy measure described above, but now for the frequency distribution of GO terms observed per representative domain.

(c) *Statistics*

All statistical analyses were carried out with the statistics software R v. 2.1.1 (R Development Core Team 2005; http://www.r-project.org/). For the principal component regression (PCR) analysis, we used the R package 'pls'.

3. RESULTS

(a) *More designable proteins show a greater capacity to produce enzymatic diversity*

Here we use two complementary measures of protein designability. The first of them is structural designability (D_S), as estimated by the length-normalized principal eigenvalue of a protein's contact density matrix (England & Shakhnovich 2003). The contact density matrix $A=(a_{ij})$ is a binary (0-1) matrix, where $a_{ij}=1$ if two non-neighbouring residues i and j ($|i-j|>1$) are in contact. The principal eigenvalue of the contact density matrix tends to be larger for proteins with more amino acid contacts per residue, adopting a value between the average number of contacts per residue and the maximal number of contacts of any given residue (Porto *et al.* 2004). The measure D_S reflects the number of groups of interacting amino acids. A large number of such groups allow more sequences to adopt a structure by relaxing energy constraints for the rest of the sequence (Shakhnovich *et al.* 2005).

Our second measure is diversity designability (D_D), which is the number of sequences from a non-redundant database (see §2) that fold into a structure, normalized by the sequence length. This second measure is vulnerable to a confounding factor, the different age of proteins. Old proteins may have more sequences associated with them than younger proteins, just because they originated early in life's evolution. To exclude this factor, we restricted our analysis of diversity designability (D_D) to a set of 112 ancient protein domains in the CATH database, which were probably present in the most recent common ancestor of all extant life (Ranea *et al.* 2006). Both measures of designability are highly correlated for this age-corrected set of domains (Spearman's $r=0.88$; $p<7.25\times10^{-53}$; table 1) and for the complete set of more than 1924 CATH domains (Spearman's $r=0.89$; $p<10^{-100}$; figure 2*a*). Similar associations have been reported for different domain datasets (Shakhnovich *et al.* 2005). They suggest that D_S is reflective of the number of sequences that adopt a structure.

We used two complementary measures of protein functional diversity. The first is a measure of diversity of enzymatic functions, based on structural similarities of enzyme active sites. The relevant information is curated in a recently developed database, which classifies enzymes into three hierarchical levels of function, of which we use the lowest (familial) level here (Pegg *et al.* 2006). We use two quantitative indicators of enzymatic functional diversity. These are FE_1, the number of enzyme families associated with a protein domain, and FE_2, which takes into account that different enzymatic functions occur at different frequencies in a set of sequences associated with a domain (see §2). We explored the association between protein designability and functional diversity for these two different notions of functional diversity.

Figure 1*a* shows an example of two structures with very different designabilities (figure 2*a*). The colour spectrum in the tertiary structure ranges from blue to red, corresponding to positions with low and high sequence diversity (D_D), respectively. The structure in figure 1*a*(i) has lower designability and lower functional diversity, as indicated by the number of associated enzymatic functions, than the structure in figure 1*a*(ii). The less designable domain is associated with two enzyme superfamilies and three families, whereas the more designable domain is associated with four enzyme superfamilies and 11 families. Figure 1*b*

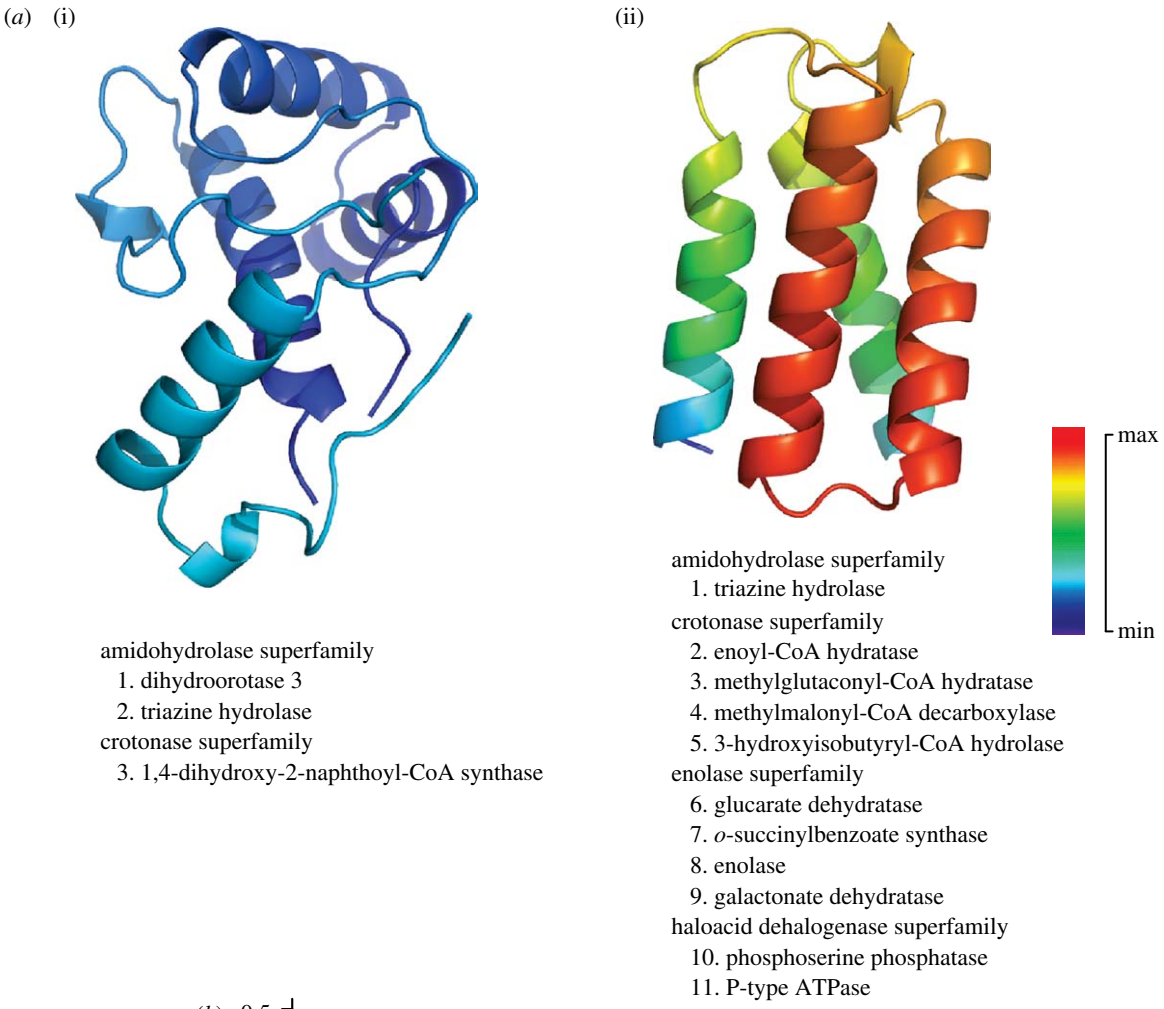


Figure 1. (*Caption opposite.*)

shows a scatterplot of D_D and enzymatic functional diversity (FE_1) for eight arbitrarily chosen ancient structures that are colour-coded in the same way. It suggests that the difference evident from figure 1*a* is not just a peculiarity of the two sequences chosen.

For the complete dataset of ancient domains, we observe a statistically significant and highly positive association between enzymatic functional diversity and D_D (Spearman's $r=0.80$; $p<2.31\times10^{-40}$; figure 3*a*). A structure with more associated sequences might be expected to have more associated functions, but this association persists if we normalize the number of functions by the total number of sequences associated with each fold (Spearman's $r=0.44$; $p<1.95\times10^{-15}$).

Figure 1. (*Opposite*.) (*a*) An example of protein domains with different designabilities and different functional diversities. For the purpose of illustration, the minimum and maximum number of sequences has been scaled linearly. Thus, the colour spectrum indicates a measure of sequence diversity, where blue (red) corresponds to minimum (maximum) sequence diversity estimated per residue. The diversity designability of a domain D_D is a domain-wide average over this sequence diversity. The enzyme families associated with each domain are listed. (i) A domain with low designability (CATH identifier: 1mw9X04: topoisomerase 1, domain 4). It has a complex fold and is associated with three enzyme families that fall into two superfamilies (Pegg *et al.* 2006). (ii) A domain with high designability (1ls1A01: the A subunit of the four-helix bundle hemerythrin domain). It has a simpler fold and is associated with 11 enzyme families and four superfamilies. Superfamilies and families are listed. (*b*) Enzymatic functional diversity (FE_1) increases with protein designability. Enzymatic functional diversity (FE_1) is expressed as the number of different enzyme families per representative CATH domain (Pegg *et al.* 2006). Eight highly conserved CATH domains (1n55A00, 1qz5A01, 1q6zA03, 1rl6A02, 1k7wA03, 1ls1A01, 1vq8V00 and 2bm0A03) have been arbitrarily chosen to illustrate the association between enzymatic functional diversity (FE_1) and designability (D_D, D_S). The Spearman rank correlation coefficient between D_S and FE_1 for these eight domains is 0.92.

We also examined the association between structural designability D_S and enzymatic functional diversity. This association is also positive, regardless of whether we normalize for the number of sequences associated with a fold (Spearman's $r=0.55$; $p<1.24\times10^{-20}$) or not (Spearman's $r=0.70$; $p<1.09\times10^{-30}$; figure 3*b*). An even higher positive association exists if we use the frequency-weighted measure of enzymatic functional diversity, FE_2 (D_D: Spearman's $r=0.88$; $p<7.13\times10^{-52}$; D_S: Spearman's $r=0.88$; $p<7.13\times10^{-52}$).

(b) *More designable proteins show greater overall diversity of molecular functions*

Our second measure of functional diversity encompasses GOAs of molecular functions. The GO database includes the most comprehensive information about functional diversity of proteins. It is not restricted to enzymes. The GO project has developed a dynamic controlled vocabulary based on three aspects of function (molecular function, process and location) that encompass complementary notions of gene functions in living cells (Ashburner *et al.* 2000). For our purpose, the appropriate aspect of function is molecular function. We used two measures of molecular functional diversity. The first (FG_1) is simply the number of molecular function annotations associated with a protein domain and the second (FG_2) weights different functions by their frequency in a set of proteins (see §2).

We observe a statistically significant and highly positive association between functional diversity (FG_1) and D_D, regardless of whether we normalize for the number of sequences per domain (Spearman's $r=0.62$; $p<1.53\times10^{-24}$) or whether we do not normalize (Spearman's $r=0.96$; $p<2.5\times10^{-79}$; figure 3*c*). We also examined the association between D_S and FG_1, which is positive independent of whether the values are normalized (Spearman's $r=0.86$; $p<1.94\times10^{-48}$) or whether they do not normalize (Spearman's $r=0.94$; $p<2.99\times10^{-68}$; figure 3*d*). An even higher positive association exists if we use the frequency-weighted measure of functional diversity, FG_2 (D_D: Spearman's $r=0.87$; $p<3.48\times10^{-50}$; D_S: Spearman's $r=0.97$; $p<4.08\times10^{-88}$).

(c) *Fold complexity and ID influence designability and diversity*

Protein designability may be correlated with a number of other protein properties. Although such properties are not the main focus of our analysis, we wanted to examine how some of them relate to functional diversity. The first of these properties is the complexity of a protein fold. Among various available measures (Arteca 1995; Enright & Leitner 2005), we use the ACO as a measure of fold complexity. ACO is the average distance on the amino acid

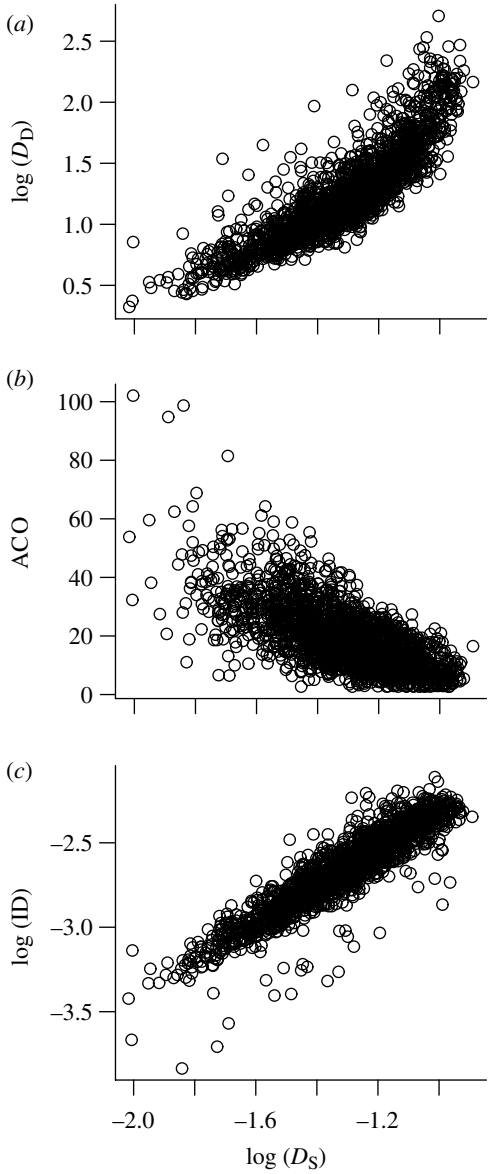


Figure 2. Designability, fold complexity and disorder are associated properties. (*a*) Diversity designability (D_D) versus structural designability (D_S). (*b*) Fold complexity (ACO) versus structural designability (D_S). (*c*) Intrinsic disorder (ID) versus structural designability (D_S). D_D corresponds to the total number of sequences per residue per representative domain. ID is calculated as a length-normalized average per representative domain. Decadic logarithm is applied.

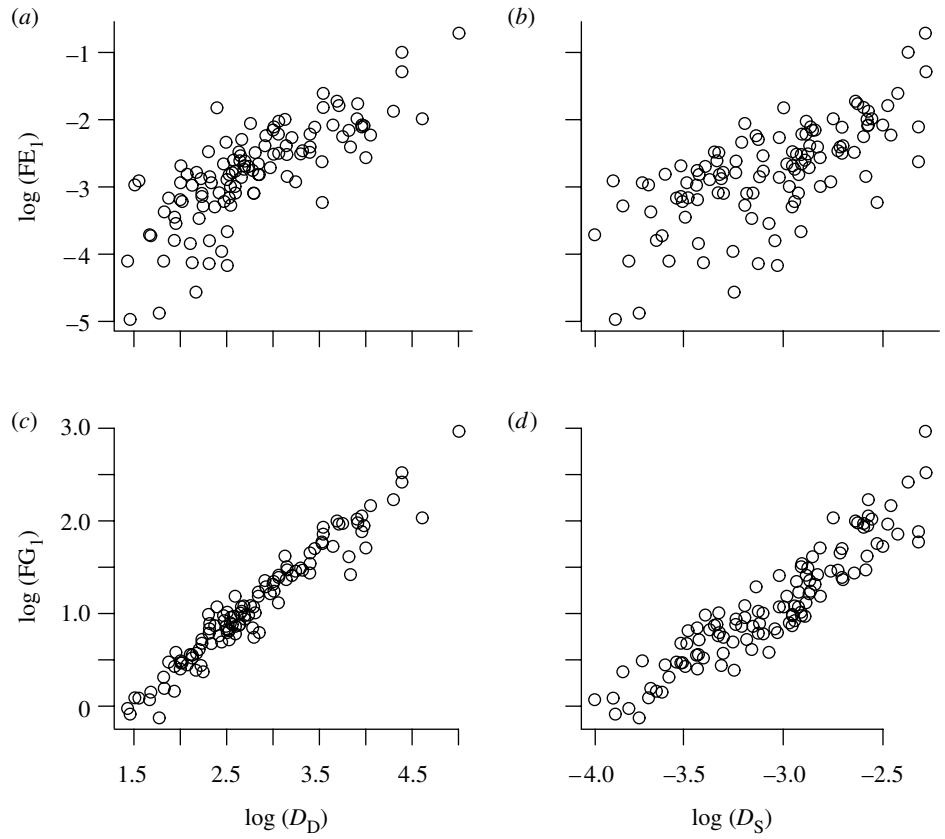


Figure 3. Functionally diverse proteins are highly designable. (*a*) Enzymatic functional diversity (FE_1) as a function of diversity designability (D_D). (*b*) Enzymatic functional diversity (FE_1) as a function of structural designability (D_S). (*c*) Molecular (gene ontology) functional diversity (FG_1) as a function of diversity designability (D_D). (*d*) Molecular (gene ontology) functional diversity (FG_1) as a function of structural designability (D_S). Functional diversity measures shown are normalized by the total number of sequences associated with each representative domain. D_D corresponds to the length-normalized number of sequences per representative domain.

sequence of two residues that contact each other in the structure. It can be thought of as a measure of how 'entangled' a structure is. It is a good predictor of a protein's folding rate, regardless of whether the folding kinetics is dominated by one or several steps (Ivankov *et al.* 2003). Highly designable proteins have low fold complexity. (D_S: Spearman's $r=-0.70$; $p<9.11\times10^{-91}$; D_D: Spearman's $r=-0.66$; $p<1.14\times10^{-27}$; figure 2*b*).

Second, we also explore the relationship between designability and a measure for the amount of conformational disorder a protein can tolerate. Highly disordered proteins are more flexible than others. The measure we use is the 'intrinsic disorder' of a protein, as defined in Dosztányi *et al.* (2005*b*). Specifically, here we use the average ID of the set of sequences associated with each CATH representative domain (see §2). We would predict that proteins with high intrinsic disorder can tolerate more sequence change, and that they might thus also be more designable. This is the case (D_S: Spearman's $r=0.92$; $p<10^{-100}$; D_D: Spearman's $r=0.90$; $p<4.07\times10^{-56}$; figure 2*c*). Not surprisingly, these properties are also associated with each other (table 1).

Because protein fold complexity and disorder are associated with designability, they might also be associated with functional diversity. This is indeed the case (table 1). The diversity of enzymatic and general molecular functions increases for short proteins (FE_1: Spearman's $r=-0.685$; $p<2.33\times10^{-29}$; FG_1: Spearman's $r=-0.94$; $p<1.22\times10^{-68}$), for proteins with low fold complexity (FE_1: Spearman's $r=-0.63$; $p<3.6\times10^{-25}$; FG_1: Spearman's $r=-0.77$; $p<2.22\times10^{-36}$) and for proteins with high intrinsic disorder (FE_1: Spearman's $r=0.71$; $p<6.22\times10^{-31}$; FG_1: Spearman's $r=0.93$; $p<1.1\times10^{-65}$).

The pairwise associations we have discussed so far may conceal subtle interactions among the multiple variables we consider here. To better disentangle their relationship, we thus performed a PCR analysis. This analysis allows us to understand how the three critical variables—designability, fold complexity and disorder—contribute to functional diversity. The results of this analysis reveal no unforeseen new relationships (figure 4). One dominant principal component accounts for more than 80% of the variance in functional diversity. This component is dominated by the positive role of designability and ID for functional diversity and by the negative role of fold complexity (figure 4). The second and third principal components contribute only 15 and 4% of the variance, respectively. Similar results (not shown) hold if diversity designability or enzyme functional diversity is used in the analysis.

4. DISCUSSION

In summary, our observations show that highly designable proteins evolve more functional innovations on large timescales. Our measures of designability estimate a given domain's ability to explore sequence space and access a

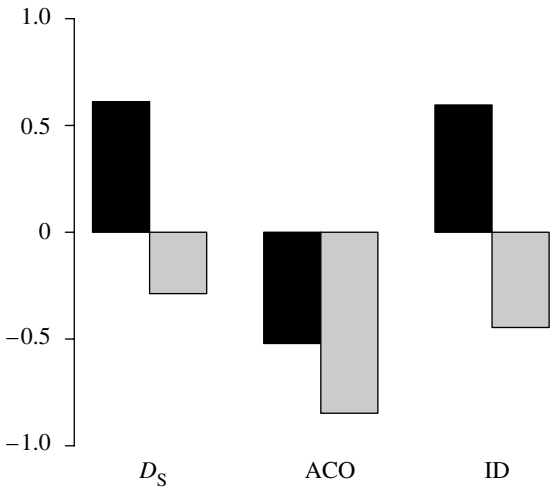


Figure 4. PCR analysis of molecular functional diversity (gene ontology) against structural designability (D_S), fold complexity (ACO) and intrinsic disorder (ID) of folds. Shown are the two principal components that together account for 96.6% of the variance observed for functional diversity. Component 1 (black bars) and component 2 (grey bars) accounts for 80.8 and 15.8% of the total variance, respectively.

diverse spectrum of functions. Because functional diversity is a record of past evolutionary innovations, this means that more designable proteins may have a greater facility to evolve new functions. In addition, because proteins of similar structure are connected in genotype space (Babajide *et al.* 1997, 2001; Bornberg-Bauer 1997; Bastolla *et al.* 1999; Wroe *et al.* 2007), more robust proteins may show greater propensity to evolve functional innovations. This association holds for two complementary measures of functional diversity: diversity of enzymatic functions and GO-based diversity of molecular functions. It also holds for two different measures of designability: one based purely on structural information and the other based on the number of sequences associated with each protein fold. The associations persist if we correct for the different numbers of sequences associated with a fold. For gene ontology annotations, these associations are also corroborated by an analysis based on a different domain dataset (Shakhnovich *et al.* 2005), whose main focus was to explain different sequence family sizes associated with different folds.

A number of other protein properties are associated with designability, and thus, not surprisingly, with functional diversity. Specifically, long proteins, proteins with complex folds (and thus proteins with slower folding rates; Ivankov *et al.* 2003) and proteins with low amounts of disorder in their tertiary structure show low functional diversity. Most of these associations have intuitive explanations. For example, it is easy to see how a high complexity of a fold may lead to smaller numbers of sequences being able to adopt a fold.

With respect to disorder in protein structures, conflicting interpretations can be brought to bear on its relationship to designability. On the one hand, a more disordered structure may be more flexible, and thus tolerate more amino acid changes, implying greater robustness and designability. On the other hand, a disordered structure may be less thermodynamically stable (Dosztányi *et al.* 2005*b*) and greater thermodynamic stability has been associated with robustness (Bastolla & Demetrius 2005; Bloom *et al.* 2006*b*). Although explanations that could resolve this conflict have been put forward (Bastolla & Demetrius 2005), such resolution is not within the scope of this contribution.

A caveat to our—and any other—comparative study is that statistical association is not equivalent to causation. Other known features (expression level, domain architecture, etc.) and unknown features of proteins may show hidden associations with functional diversity that may explain some of its variation. To identify such features would be a worthwhile subject of future studies, as would be the reduction of biases in the data, as well as the elimination of errors contained in some measures of structural differences among proteins. For example, the ID estimate we use (IUPred) has a true positive rate of 85% (Dosztányi *et al.* 2005*b*), which could be improved.

Complex relationships with other variables notwithstanding, it is clear that designable and robust proteins have evolved many novel functions. This shows that a pattern derived from recent experimental findings, and applicable only to laboratory time-scales, also holds on vastly greater geological time-scales (Aharoni *et al.* 2005; Bloom *et al.* 2006*b*). The possible explanation has its root in how populations explore vast sequence spaces: populations of highly robust folds can explore sequence space rapidly, and thus access large amounts of structural diversity in their neighbourhood (Wagner 2007). A small fraction of this diversity can subsequently give rise to proteins with new functions.

A.W. acknowledges support through grant 315200-116814 from the Swiss National Foundation, as well as support from the Santa Fe Institute.

REFERENCES

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C. & Tawfik, D. S. 2005 The evolvability of promiscuous protein functions. *Nat. Genet.* **37**, 73–76. (doi:10.1038/ng1482)

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)

Amitai, G., Gupta, R. D. & Tawfik, D. S. 2007 Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67–78. (doi:10.2976/1.2739115)

Arteca, G. 1995 Scaling regimes of molecular size and self-entanglements in very compact proteins. *Phys. Rev. Lett.* **51**, 2600–2610. (doi:10.1103/PhysRevE.51.2600)

Ashburner, M. *et al.* 2000 Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)

Babajide, A., Hofacker, I. L., Sippl, M. J. & Stadler, P. F. 1997 Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des.* **2**, 261–269. (doi:10.1016/S1359-0278(97)00037-0)

Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A. S. & Stadler, P. F. 2001 Exploring protein sequence space using knowledge-based potentials. *J. Theor. Biol.* **212**, 35–46. (doi:10.1006/jtbi.2001.2343)

Bastolla, U. & Demetrius, L. 2005 Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng. Des. Sel.* **18**, 405–415. (doi:10.1093/protein/gzi045)

Bastolla, U., Roman, H. E. & Vendruscolo, M. 1999 Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**, 49–64. (doi:10.1006/jtbi.1999.0975)

Bastolla, U., Porto, M., Eduardo Roman, M. H. & Vendruscolo, M. H. 2003 Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J. Mol. Evol.* **56**, 243–254. (doi:10.1007/s00239-002-2350-0)

Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. 2006*a* Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761. (doi:10.1093/molbev/msl040)

Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. 2006*b* Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874. (doi:10.1073/pnas.0510098103)

Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. 2007 Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Dir.* **2**, 17. (doi:10.1186/1745-6150-2-17)

Bornberg-Bauer, E. 1997 How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403.

Camon, E. *et al.* 2004 The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res.* **32**, D262–D266. (doi:10.1093/nar/gkh021)

Chothia, C. & Lesk, A. M. 1986 The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. 2005*a* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434. (doi:10.1093/bioinformatics/bti541)

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. 2005*b* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839. (doi:10.1016/j.jmb.2005.01.071)

England, J. L. & Shakhnovich, E. I. 2003 Structural determinant of protein designability. *Phys. Rev. Lett.* **90**, 218101. (doi:10.1103/PhysRevLett.90.218101)

Enright, M. B. & Leitner, D. M. 2005 Mass fractal dimension and the compactness of proteins. *Phys. Rev. E* **71**, 011912. (doi:10.1103/PhysRevE.71.011912)

Greene, L. H. *et al.* 2007 The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291–D297. (doi:10.1093/nar/gkl959)

Holm, L. & Sander, C. 1998 Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423–429. (doi:10.1093/bioinformatics/14.5.423)

Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. 2003 Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062. (doi:10.1110/ps.0302503)

Li, H., Helling, R., Tang, C. & Wingreen, N. 1996 Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669. (doi:10.1126/science.273.5275.666)

Müller, G. B. & Wagner, G. P. 1991 Novelty in evolution: restructuring the concept. *Annu. Rev. Ecol. Syst.* **22**, 229–256. (doi:10.1146/annurev.es.22.110191.001305)

Pegg, S. C. *et al.* 2006 Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **45**, 2545–2555. (doi:10.1021/bi052101l)

Plaxco, K. W., Simons, K. T. & Baker, D. 1998 Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994. (doi:10.1006/jmbi.1998.1645)

Porto, M., Bastolla, U., Roman, H. E. & Vendruscolo, M. 2004 Reconstruction of protein structures from a vectorial representation. *Phys. Rev. Lett.* **92**, 218101. (doi:10.1103/PhysRevLett.92.218101)

Ranea, J. A., Sillero, A., Thornton, J. M. & Orengo, C. A. 2006 Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.* **63**, 513–525. (doi:10.1007/s00239-005-0289-7)

Schultes, E. A. & Bartel, D. P. 2000 One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448–452. (doi:10.1126/science.289.5478.448)

Shakhnovich, B. E., Deeds, E., Delisi, C. & Shakhnovich, E. 2005 Protein structure and evolutionary history determine sequence space topology. *Genome Res.* **15**, 385–392. (doi:10.1101/gr.3133605)

The UniProt, C. 2007 The universal protein resource (UniProt). *Nucleic Acids Res.* **35**, D193–D197. (doi:10.1093/nar/gkl929)

Wagner, A. 2007 Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B* **275**, 91–100. (doi:10.1098/rspb.2007.1137)

Wroe, R., Chan, H. S. & Bornberg-Bauer, E. 2007 A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J.* **1**, 79–87. (doi:10.2976/1.2739116)

4. Evolutionary innovations and the organization of protein functions in genotype space

Ferrada E and Wagner A. Evolutionary Innovations and the Organization of Protein Functions in Genotype Space. *PLoS ONE* 5: e14172 (2010).

OPEN ACCESS Freely available online

PLoS one

Evolutionary Innovations and the Organization of Protein Functions in Genotype Space

Evandro Ferrada[1,3]*, **Andreas Wagner**[1,2,3,4]

1 Department of Biochemistry, University of Zurich, Zurich, Switzerland, **2** The Santa Fe Institute, Santa Fe, New Mexico, United States of America, **3** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **4** Department of Biology, University of New Mexico, Albuquerque, New Mexico, United States of America

Abstract

The organization of protein structures in protein genotype space is well studied. The same does not hold for protein functions, whose organization is important to understand how novel protein functions can arise through blind evolutionary searches of sequence space. In systems other than proteins, two organizational features of genotype space facilitate phenotypic innovation. The first is that genotypes with the same phenotype form vast and connected genotype networks. The second is that different neighborhoods in this space contain different novel phenotypes. We here characterize the organization of enzymatic functions in protein genotype space, using a data set of more than 30,000 proteins with known structure and function. We show that different neighborhoods of genotype space contain proteins with very different functions. This property both facilitates evolutionary innovation through exploration of a genotype network, and it constrains the evolution of novel phenotypes. The phenotypic diversity of different neighborhoods is caused by the fact that some functions can be carried out by multiple structures. We show that the space of protein functions is not homogeneous, and different genotype neighborhoods tend to contain a different spectrum of functions, whose diversity increases with increasing distance of these neighborhoods in sequence space. Whether a protein with a given function can evolve specific new functions is thus determined by the protein's location in sequence space.

Citation: Ferrada E, Wagner A (2010) Evolutionary Innovations and the Organization of Protein Functions in Genotype Space. PLoS ONE 5(11): e14172. doi:10.1371/journal.pone.0014172

Editor: Johannes Jaeger, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Spain

Received July 21, 2010; **Accepted** October 28, 2010; **Published** November 30, 2010

Funding: AW acknowledges support through Swiss National Science Foundation grants 315200-116814, 315200-119697 and 315230-129708, as well as through the YeastX project of SystemsX.ch, and the University Priority Research Program in systems biology at the University of Zurich. EF acknowledges support through UZH Forschungskredit. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: e.ferrada@bioc.uzh.ch

Introduction

During more than half a century of protein research, an enormous amount of data about protein sequences, their structures, and their functions has accumulated. To organize the vast number of known protein sequences, the concept of a sequence space is useful [1]. Two sequences in this space have a distance, which can be measured in various ways [2,3]. The simplest such measure is the sequence distance, the number or percentage of amino acid changes needed to transform one protein onto another. Two sequences in this space can have either the same or a different fold. This fold is the three-dimensional arrangement of their amino acids, and typically involves a specific arrangement of α-helices and/or β-sheets, the secondary structure elements of proteins. The organization of protein structures in sequence space has several general features.

First, only a small fraction of protein sequences, perhaps no larger than 10^{-4}, may adopt a stable, well-defined structure [4]. Considering the astronomical size of sequence space, however, this still leaves many proteins that fold. For example, for proteins of length 100 amino acids, sequence space has 20^{100} members. Even if only one in 10^4 of them adopts a stable structure, approximately 10^{126} foldable sequences exist in this space.

Second, the existing repertoire of protein folds is small [5,6], and the number of sequences greatly surpasses its size.

Third, many of a protein's immediate neighbors – sequences differing from it in a single amino acid – typically have the same fold as the protein itself [7–9].

Fourth, even very distant sequences can have the same fold [10,11]. If two such sequences have the same common ancestor, they are often referred to as members of the same *protein family* [6]. Such unambiguous common ancestry can usually be identified for sequences that differ in up to 60 to 70 percent of their amino acids [12]. Two sequences in the same family can be connected through a series of amino acid changes that traverse a fraction of sequence space while leaving the structure unchanged. When common ancestry can be claimed based on criteria such as common aspects of structure or function, families of proteins are grouped into superfamilies. Superfamilies share a common fold and diverge on average around 70 to 80 percent in sequence space. Sets of superfamilies that share the same three-dimensional arrangement of secondary structure are grouped into the same fold. Amino acid sequences with the same fold can be very different. Based on a systematic comparison of many divergent sequences with shared folds, Rost [11] observed that such sequences can have more than 95 percent divergence.

Fifth, the number of sequences per fold may vary widely. For example, mutagenesis experiments suggest that the amino acid sequences forming an enzyme with the same structure and function as chorismate mutase may occupy a fraction 10^{-23} of

sequence space [13], whereas sequences forming a functional β-lactamase domain occupy merely one 10^{-64th} of sequence space [14]. Structures adopted by many sequences are commonly called highly designable [15,16]. There has been increasing interest in highly designable proteins due to their use as 'scaffolds' in the design of new protein functions [17]. One remarkable example is the zinc finger domain, which is robust to point mutations in alanine scanning experiments [18], and has proven useful in designing new DNA binding proteins [19].

Taken together, these observations suggest that the protein sequences adopting the same structure form connected networks of sequences that can reach far through sequences space and that have varying size. These properties are not only observed for real proteins, but also for lattice proteins, and other generic models of protein folding [15,20–23]. They emerge from generic physico-chemical properties of the protein folding process. In other words, they are characteristic of the mapping between genotypes (sequences) and phenotypes (structures) that exists for proteins. We will call a connected network of sequences with the same structure a *genotype network*.

Similar to information about protein structures, which is abundant, thousands of proteins have known and well-characterized *functions*. However, while several authors studied the distribution of structures in sequence space [22,24–25], we know much less about how functions are distributed through sequence space. This question is the main focus of our work.

The need to assign a function to newly identified protein sequences has driven research into the conservation of protein functions as sequences diverge. Several studies using methods of sequence comparison agree that functional conservation is common if two proteins possess more than 50% sequence identity [26–30]. For gene ontology functional annotations, more than 90 percent of protein pairs over 50% sequence identity have the same function [31]. However, a study dissenting from the conclusion of earlier work found that fewer than 30 percent of proteins with more than 50 percent sequence identity have identical enzymatic functions [32].

Information like this makes it clear that we cannot simply extrapolate from structure to function. To be sure, some proteins, such as oxygen-binding globins have the same structure and function, despite great sequence divergence [10]. However, other proteins have the same structure but different functions. Examples include proteins with the TIM-barrel fold, which is associated with many enzymatic functions [33]. In addition, many functions can be carried out by proteins with different structures. Examples include DNA polymerases, which use similar catalytic mechanisms, but diverse structures, to replicate DNA [34].

Taken together, these observations show that the relationship between sequence, structure, and function is complex. Thus, any analysis aiming to understand the organization of protein functions in sequence space must not tie itself too closely to protein structure, while respecting that structure constrains function. The biggest obstacle to such an analysis is to describe and categorize protein functions for many proteins. We circumvent this obstacle by focusing on enzymes, proteins for which a well-established, albeit imperfect, functional classification exists.

To understand how protein functions are organized in sequence space is important for at least three reasons. First, it may help guide the development of methods for protein function annotation (which is not our focus here). Second, it may help identify functions that can be performed by a large number of sequences. Experimental evidence suggests that different functions may differ by orders of magnitude in the numbers of proteins that perform them [13,14,35], hinting that protein functions may differ in their designability just like structures do. Being able to distinguish functions that are adopted by many proteins from those adopted by few proteins would help identify functions that are easily created or modified through directed evolution experiments and rational protein engineering. Third, and most important, it may shed light on one of the key unsolved problems in evolutionary biology, namely how new functions arise in evolution. Proteins are ideal systems for systematic studies of biological systems' ability to innovate. The reason is that we already have so much information about them.

In a variety of biological systems, the existence of extended genotype networks facilitates the evolution of novel phenotypes [36–38]. The reason is that different regions of genotype space contain different kinds of new phenotypes. Such phenotypes can be encountered through (neutral) exploration of a genotype network and its neighborhood in sequence space. We do not know whether the same holds for proteins, that is, whether different regions of protein genotype space contain proteins with different novel functions.

To address the issues we just discussed, we use a large dataset of protein sequences with known function and structure. Our analysis uses the concept of a protein's neighborhood in sequence space, a region comprising all sequences up to some maximal distance from the protein. We show that different neighborhoods in protein sequence space contain different functions. We discuss the implications of this observation, the limitations of our procedure, and propose a general perspective on the organization of protein functions in sequence space.

Methods

Protein sequences. Structural and functional annotation

We obtained protein sequences from Uniprot [39]. Specifically, we used the dataset compiled in UniProtKB/Swiss-Prot that corresponds to manually curated protein sequences. By September 2009, this dataset was composed of 495,880 sequences for which experimental details and computed features were available. To facilitate protein comparison, we restricted our study to single domain proteins longer than 50 amino acids. The structural information we used is based on the CATH classification of protein structure domains (v.3.2.0) [40]. Throughout, we use the concepts of structure and domain interchangeably and define it at the level of homologous superfamily.

We mapped domains to Uniprot sequences using HMM libraries from CATH and the software HMMER [41], assigning domains to sequences at an e-value of 0.001. Using this procedure, we found a total of 174,853 single domain sequences. Because we aimed at a broad characterization of sequence space, we did not filter our dataset for redundant sequences, but simply restricted the allowed sequence identity between pairs of sequences to at most 99 percent, thus obtaining a dataset of 136,677 sequences. We discarded sequences tagged with any of the keywords: ''putative'', ''probable'', ''by homology''. As a source of functional annotation, we used the Enzyme Nomenclature Database (EC) [42]. Since the EC classification distinguishes four different hierarchical levels of enzyme function, we used only EC assignations that possess numerical descriptors for all of the 4 levels of the hierarchy. Using information in this database, we arrived at our final data set, which comprises 39,529 protein sequences. These sequences correspond to 1,343 enzyme types classified under the EC system. They adopt 457 different structures, as indicated by their CATH domains.

Our next goal was to align sequences in our data set, in order to estimate their pairwise distance in sequence space. To do so, we grouped our sequences according to the CATH domains they had. For each sequence, we kept only the regions for which HMM

profiles had detected significant sequence similarity between sequences. This procedure discards uninformative regions of proteins and improves the quality of the subsequent alignments, which we carried out with ClustalW [43]. We also tested the performance of structural alignments using T-coffee [44] and found that in the case of our dataset, Clustalw and T-coffee produced similar results. The number of sequences per multiple sequence alignment varied according to domains, with a median of 12 sequences per alignment. For further analyses we included only proteins where, after multiple sequence alignment, at most 10 percent of positions were gaps, and no more than 10 percent of any one amino acids sequence contained gaps.

We carried out two different analyses of our data. First, we characterized, for proteins with a given structure, how their functions were distributed across sequence space. To this end, we focused on 36 different structures for which at least 10 sequences are known. Specifically, these structures have between 10 and 4,132 associated sequences. Except for the TIM barrel, we carried these analyses out exhaustively, that is, considering all possible pairwise comparisons between sequences that share a structure domain (see figure legends for details). Second, we examined the distribution of functions regardless of the structures performing them. In this analysis, a complication is that proteins with different structures can have different lengths. To facilitate their embedding in the same genotype space, we focused only on alignments with sequences no shorter than 100 amino acids. The resulting (reduced) data set had 28,862 sequences, 337 different structures, and 1,036 enzyme functions. We then selected random sections of 100 residues from each multiple sequence alignment, calculated the desired statistic from the resulting resampled data, and repeated this resampling and calculation procedure a total of 10 times. (Since proteins with more than 10 percent of gaps are discarded, each one of the 10 samples comprises on average 28,862 sequences, 337 different structures, and 1,036 enzyme functions.) We performed the neighborhood analysis described below on each of these 10 samples, and report results as means and standard deviations over these 10 samples.

Results

To characterize the distribution of protein functions in sequence space, we used a comprehensive protein dataset of 39,529 sequences that adopt 457 single-domain structures. In the following, we refer to them simply as structures. The functions we consider are based on the enzyme commission (EC) [42] classification, which distinguishes four different hierarchical levels of enzyme function. The top level comprises six enzyme classes, namely oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. Each class is subdivided into three further hierarchical levels whose interpretation differs among classes. In this classification system, individual enzymes are assigned a four-digit number where each digit reveals increasing details about enzyme function. For example, the enzyme tryptophan synthase with EC number 4.2.1.20 is a lyase that catalyzes the conversion of indole and serine to tryptophan. Although the EC classification has well-known limitations (eg. see [30]), it is the best-established and most widely used system for classifying enzymes, which are the most prominent protein class. (By March 2010, 57 percent of proteins in the Protein Data Bank [45], a repository of protein structure information, have at least one enzymatic function). For our data set, the bottom, finest-grained level of this classification comprises 1,343 different enzymes. For this data set, Figure S1a shows the distribution of the number of sequences per structure, and Figure S1b shows the number of sequences per function.

Although our data set may seem enormous, we note that it still represents a very sparse sampling of sequence space. For example, approximately 60 percent of functions are represented by fewer than 10 sequences per function. Also, two proteins with the same structure and/or function in our data are typically highly divergent, with a median amino acid divergence of no less than 55 percent (Figure S2a and S2b).

Most enzymatic functions are associated with few structures

Any given function in our data set may be carried out by proteins with only one structure, or by multiple different structures. We call the latter kind of function *structurally promiscuous*, because it is not tied to any one structure. Figure 1a shows a histogram of the number of structures associated with a function

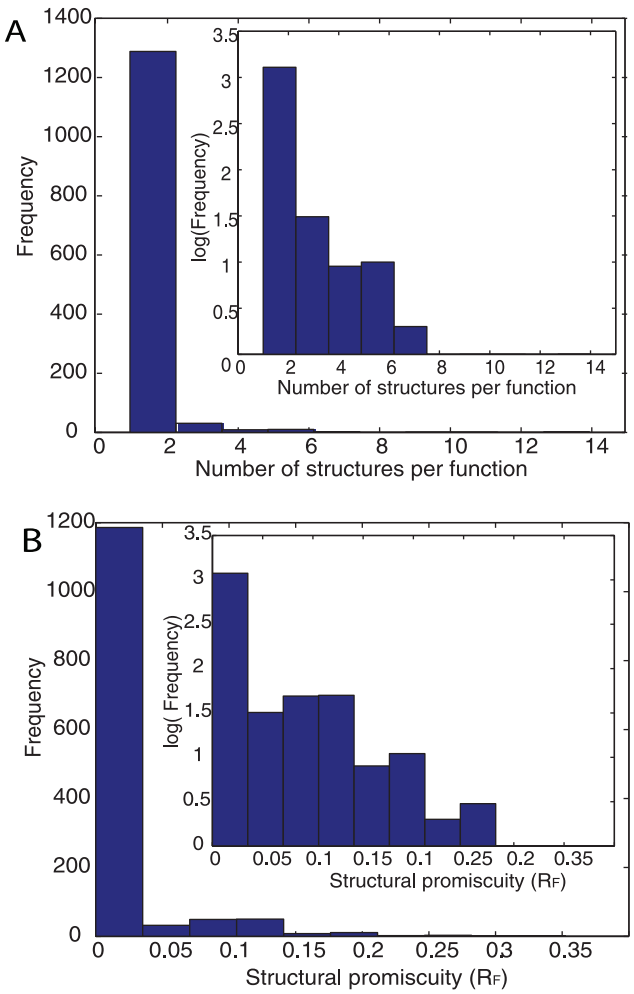


Figure 1. Distribution of structures over functions. (**a**) *Distribution of the number of structures associated with a particular function.* The total number of different structures (457) in our dataset composed of 39.529 sequences are classified according to the enzyme function that they perform and counted (min = 1 ; max = 14 ; mean = 1.2). The inset shows the same distribution, but with a $\log_{10}$-transformed vertical axis. (**b**) *Distribution of structural promiscuity.* Structural promiscuity (R_F) is an entropy-like measure (see main text) calculated from the distribution of enzyme functions over different protein domains. The data shown is based on the finest-grained, fourth level of the EC hierarchy. (min = 0.0; max = 0.35; mean = 0.01).
doi:10.1371/journal.pone.0014172.g001

for the 1,343 lowest level enzymatic functions we discuss here. This distribution is highly skewed, with 86 percent of the functions carried out only by one structure and three maximally promiscuous functions carried out by 9, 11 and 14 structures, respectively. These functions are RNA polymerase (EC = 2.7.7.6); cytochrome oxidase (EC = 1.9.3.1) and DNA polymerase (EC = 2.7.7.7). Figure S3 shows that the distribution remains skewed if we control for the number of sequences known per structure.

We next extended previous work [30] by defining a measure R_F of the promiscuity of any given function. We focus on only those sequences that perform a given function *F*. For any given protein structure *i* (out of *N* total structures), we denote as *f(i)* the fraction of sequences among all proteins that perform the function F and fold into structure *i*. The sum of the *f(i)*'s over all structures will add to one. The Shannon entropy of the distribution of the non-zero *f(i)*'s is given by $-\sum_{i=1,f(i)\neq 0}^{i=N} f(i)\ln f(i))$, where ln denotes the natural logarithm. The maximal value of this entropy is ln N, which is attained if every structure is equally likely to perform the function F. Its minimal value of zero is reached if the function is carried out by only one domain k, such that $f(k) = 1$ and all other $f(i) = 0$. These observations motivate the definition of structural promiscuity as $R_F = [-\sum_{i=1,f(i)\neq 0}^{i=N} f(i)\ln f(i)]/\ln N$, which is an entropy normalized to the interval zero (low promiscuity) and 1 (highest promiscuity). R_F adopts its minimum for functions associated only with a single structure. It would attain a maximum for a function that is equally likely to be performed by any structure. (Such a function may not exist.) Figure 1b shows the distribution of R_F. This distribution is again highly skewed, with a minimum of 0 for 1,161 (86 percent) of functions that are executed only by single domains. The maximal value observed is 0.35. This highest value is attained by DNA-polymerases (EC.2.7.7.7), which are well known to be structurally diverse [46]. It is followed by type II restriction enzymes (rank 2) and ubiquitin carboxyl-terminal hydrolases (rank 3). Table 1 shows the ten most structurally promiscuous enzyme functions. We note that this measure of promiscuity R_F weights different structures according to the fraction of known sequences adopting them. It can thus give different results from simpler measures based on counting the number of sequences or structures per function.

Table 1. The ten most structurally promiscuous functions.

	EC number	N structures	*R_F	Catalytic activity
1	EC = 2.7.7.7	14	0.35	DNA-directed DNA polymerase.
2	EC = 3.1.21.4	7	0.29	Type II site-specific deoxyribonuclease
3	EC = 3.1.2.15	6	0.26	Ubiquitin thiolesterase.
4	EC = 1.6.5.3	6	0.26	NADH dehydrogenase (ubiquinone).
5	EC = 2.7.7.48	6	0.25	RNA-directed RNA polymerase.
6	EC = 2.7.7.49	5	0.22	RNA-directed DNA polymerase.
7	EC = 1.14.13.39	4	0.22	4-hydroxyphenylacetate 3-monooxygenase.
8	EC = 3.1.3.2	6	0.21	Acid phosphatase.
9	EC = 2.5.1.18	4	0.20	Glutathione transferase.
10	EC = 2.7.7.6	9	0.20	DNA-directed RNA polymerase.

*(R_F). Structural promiscuity. (See main text).
doi:10.1371/journal.pone.0014172.t001

The distributions we just presented may reflect underlying properties of sequence space, but also results of biases in existing knowledge about different structures or functions. The most obvious such bias comes from the extent to which different structures and functions have been characterized. It is reflected in the different numbers of sequences that are known for them. Figure S4a and S4b shows that this amount of information can affect estimates of the structural promiscuity of a given function. The figure demonstrates that both the number of structures known to carry out a given function, and the structural promiscuity of a function increase with the number of sequences that are associated with the function. These observations suggest that low structural promiscuity of a function may be more apparent than real, and that promiscuity will increase as more proteins with a given function become characterized.

To summarize our analysis so far, relatively few functions are carried out by multiple structures, but this number would increase as more protein sequences will become characterized. In the supplementary material (File S1), we extend this analysis to the highest level of the EC hierarchy (Figures S5, S6, S7, S8, S9), where we observe similar patterns. In addition, extending previous work [30], we also analyze the distribution of the number of functions per structure (Figures S7). This distribution is similarly skewed, with most structures having single functions, and a minority of structures adopting multiple functions.

Phenotype neighborhoods

Thus far, we have examined global aspects of the organization of enzymatic functions, disregarding where the proteins carrying out these functions occur in sequence space. We next turn to a more local analysis that focuses on different neighborhoods of sequence space. We define a neighborhood $N_G(r)$ of a protein sequence (genotype) *G*, as the set of sequences that differ in no more than a number or percentage *r* of its amino acids from *G* itself. Put differently, a neighborhood $N_G(r)$ is a ball of radius *r* around *G*. With this notion in hand, we ask whether different neighborhoods differ in the kinds of functions they contain. That is, consider two protein sequences G_1 and G_2 with sequence distance *d*, and the neighborhoods $N_{G1}(r)$ and $N_{G2}(r)$ around them (with some given radius *r*) (Figure 2). The neighborhood of G_1, $N_{G1}(r)$ contains sequences that carry out some set S_1 of enzymatic

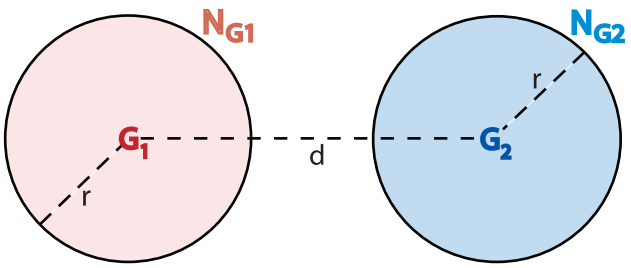


Figure 2. Genotype neighborhoods. Illustration of genotype neighborhoods by a schematic two-dimensional projection of protein sequence space. The neighborhood of a genotype ($N_{G1}(r)$) is defined as the set of all the genotypes found at a sequence distance equal or shorter than a radius (r) from the genotype of interest. Two such neighborhoods may contain different sets of functions, S_1 and S_2, respectively. We define the fraction of functions unique to a neighborhood as $F_u := (|S_1|+|S_2|-2|S_1 \cap S_2|)/|S_1 \cup S_2|$.
doi:10.1371/journal.pone.0014172.g002

functions. Similarly, $N_{G2}(r)$ contains sequences that carry out some set S_2 of enzymatic functions. The number of functions that occur in both neighborhoods equals $|S_1 \cap S_2|$, where $|X|$ denotes the number of elements in a set X. The set of all functions that are found in at least one of the two neighborhoods is $(S_1 \cup S_2)$. We define the fraction of functions that occur in the neighborhoods of one but not the other sequence as $F_u := (|S_1|+|S_2|-2|S_1 \cap S_2|)/|S_1 \cup S_2|$. For brevity, we will refer to it as the fraction of functions unique to a neighborhood. This does not mean that these functions occur nowhere else in sequence space. They just do not occur in the other neighborhood examined. F_u depends on the distance d between G_1 and G_2 and on the neighborhood radius r. We explore this dependency below.

Different genotypic neighborhoods contain highly diverse functions

Figure 3a shows a heat-map of the fraction F_u of functions unique to a sequence neighborhood, for our entire data set, and for sequences G_1 and G_2 whose distances d vary, as well as for sequence neighborhoods of various sizes r (smaller than d). The region where the two neighborhoods do not overlap, that is, where $r<d/2$, is indicated in the figure by a dashed line. For the data in this figure, we chose the neighborhood centers G_1 and G_2 regardless of the structure and function of G_1 and G_2. Perhaps of the greatest interest are neighborhoods with small radius r. They contain functions that can be reached via a small number of changes from its center G_i.

Two general observations emerge from the figure. First, at any neighborhood size r, the fraction of unique functions increases rapidly with the distance between the neighborhood centers G_1 and G_2. For a select number of sizes r, this relationship is shown also in Figure 3b, which displays F_u as a fraction of the sequence distance between G_1 and G_2. (The large standard deviations of the data at low values of d reflect the very sparse sampling of sequence space at low d.) For example, if two different sequences G_1 and G_2 of length 100 amino acids differ at only 20 percent of their amino acids, their respective neighborhoods of radius five (which correspond to sequences differing from them in no more than five percent of their amino acids) have merely 50 percent of their functions in common (Figure 3b). In other words, fifty percent of these functions are reachable from one sequence (by no more than five amino acid changes), but not from the other. More generally, small neighborhoods of two distant proteins will generally contain very different functions.

The second general feature occurs at distances between G_1 and G_2 that exceed d = 80. Here, the fraction of unique functions F_u rapidly increases to a value close to one, regardless of the neighborhood radius. This means that neighborhoods that are very far apart in sequence space contain mostly different functions. We explain below that this feature arises from the fact that highly dissimilar proteins with the same structure, proteins that are not from the same family (d larger than 80 percent) generally have different functions.

Different genotypic neighborhoods of proteins with a given structure contain highly diverse functions

The previous analysis focused on the distribution of functions in different sequence space neighborhoods, regardless of the structure or function of the proteins G1 and G2 in the neighborhood centers (Figure 2). We next asked whether similar distributions also exist if G_1 and G_2 (Figure 2) have the same structure. This is of course only possible for structures for which many sequences are available. The structure with most associated sequences in our dataset is the TIM barrel. It is represented by 4,132 sequences. These 4,132 sequences carry out 53 different enzymatic functions that cover 5 out of the 6 EC major classes and are widely spread through sequences space (Figure S10). Figure 4a shows, analogous to our analysis above, the fraction of unique enzyme functions (F_u) found in pairwise comparisons of different neighborhoods in sequence space, when considering only sequences known to fold into the TIM barrel domain. The qualitative features we observed above are also present for the TIM barrel domain. First, the fraction of unique functions increases with increasing sequence

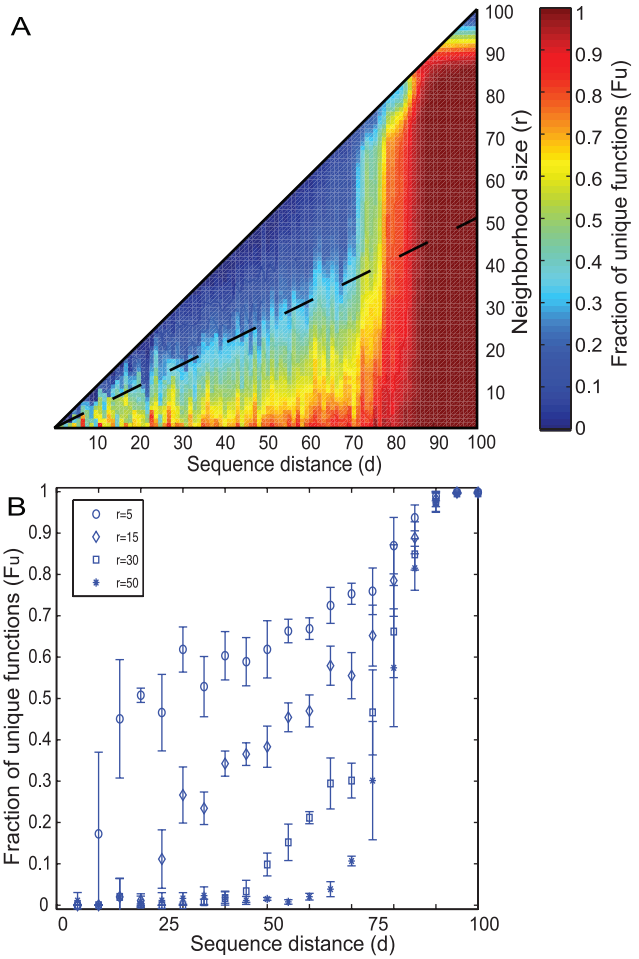


Figure 3. Different genotypic neighborhoods contain highly diverse functions. (**a**) The figure shows a heatmap of the fraction of unique functions (F_u) at different combinations of neighborhood radii (r) and sequences distances (d). The dataset analyzed here is based on 10 random subsets of 28,862 sequences from our original data, where we required that each sequence in each subset is longer than 100 amino acids. (The sequences in each subset adopted, on average 337 structures and perform 1,036 different enzyme functions.) From each of these 10 subsets, we then chose 10^5 pairs of sequences at random, and computed their values of r, d, and F_u. We repeated this random selection of 10^5 sequence pairs n times, until the results no longer changed. For the dataset of the figure, this convergence occurred around $n = 10$, but data are shown for $n = 100$. The heatmap shows the average values across the 10 samples observed for each combination of distance and radius. (**b**) Fraction of unique functions F_u versus sequence distance (expressed in percent) at constant neighborhood radii, as shown in the legend. Due to the sparsity of data, we grouped values into 20 different distance bins, each spanning d = 5. Error bars represent standard errors calculated for each of these 20 bins.
doi:10.1371/journal.pone.0014172.g003

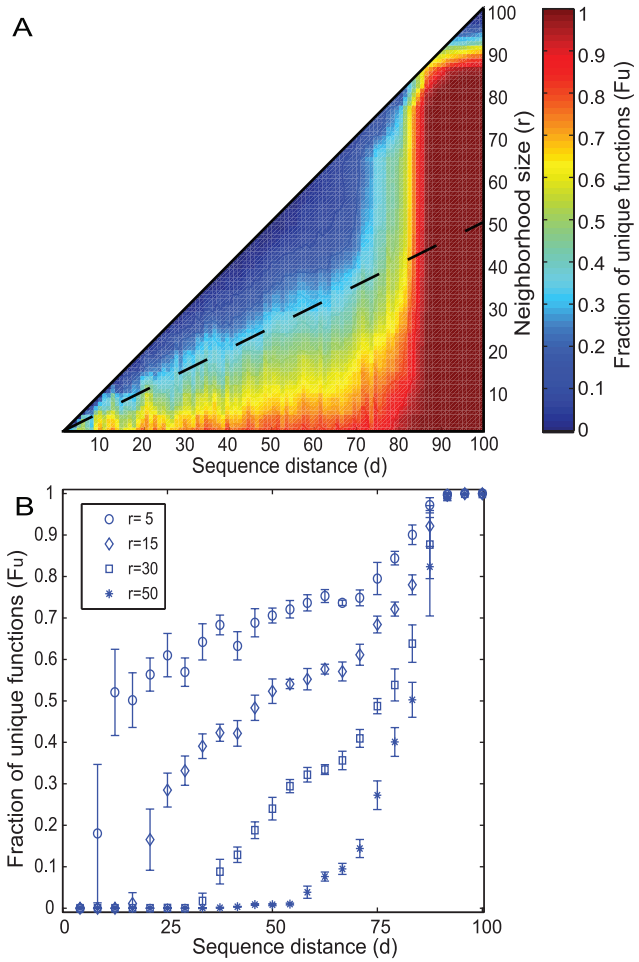


Figure 4. Genotypic neighborhoods of the TIM barrel domain. The figure shows the dependency between the radius and distance of genotype neighborhoods, and the fraction F_u of functions unique to one neighborhood, for sequences adopting the TIM barrel domain (see Methods). (**a**) Heatmap of the fraction of unique functions (F_u) at different combinations of neighborhood radii (*r*) and sequences distances (*d*). We analysed these 4,132 sequences exhaustively. That is, for all possible pairwise sequence comparisons we computed their values of r, d and F_u. The heatmap shows values of F_u at each combination of *d* and *r*. (**b**) Fraction of unique functions versus sequence distance (expressed in percent) at constant neighborhood radii, as shown in the legend. Due to the sparsity of data, we grouped values into 20 different distance bins, each spanning d = 5. Error bars represent standard errors calculated for each of these 20 bins.
doi:10.1371/journal.pone.0014172.g004

distance of the neighborhood centers G_1 and G_2 (Figure 4b). Second, at large distances of G_1 and G_2, most functions are unique, regardless of the neighborhood radius *r*.

To exclude the possibility that these observations are peculiarities of the TIM barrel domain, we carried out independent analyses for those 36 structures for which the most sequences were available. Together, they comprise a total of 18,117 sequences with lengths ranging from 100 to 400 amino acids, and span 434 enzymatic functions covering all 6 EC classes. In lieu of presenting 36 plots, figure S11 shows data averaged over all 36 structures. Its panels show the fraction F_u of unique functions and how it depends on sequence distance *d* and neighborhood radius *r*, exactly as for Figures 3a and 3b. Distances and radii are shown as percentages of total protein length. The figure shows that these 36 structures have properties qualitatively similar to that of the TIM barrel, except that the dramatic increase in F_u occurs over a broader range of sequence distances *d* (between ca. 70 and 90 percent, Figure 3a). This observation can be explained if different structures differ in the divergence that two sequences encoding them typically have. Figure S3b shows that this is indeed the case. It is based on the 337 structures that have more than one sequence in our data, and shows that the divergence of these sequences varies broadly around a large median of 92 percent. (For the TIM barrel domain, the maximal distance among sequences is 100%.)

Neighborhood diversity in functions depends on functionally versatile protein families

Thus far, we saw that the fraction of unique phenotypes increases with increasing distance of two genotypic neighborhoods, regardless of whether these neighborhoods center on proteins with the same structure (Figures 3 and 4) or on proteins with different structure (Figure S11). Our next analysis shows that this high neighborhood diversity comes from the fact that proteins in a given protein family can have multiple functions. Recall that a protein family, as used here, is a set of proteins with the same structure, and a sequence distance lower than 70 percent. Figure S12 shows that the sequences adopting any one structure often fall into multiple families.

If neighborhood diversity depends on functional diversity of proteins in the same family, then an analysis of this diversity, but for a subset of protein families with only one function per family should lead to a fundamentally different result from that observed in Figures 3, 4, and S11. We thus repeated our analysis of functional diversity for the TIM barrel structure, but for a subset of its protein families that carry out only single functions (Figure S13). The analysis shows that different neighborhoods now contain identical functions for all neighborhood centers with less than d = 80 percent divergence, which is the divergence of these TIM barrel families. Functional diversity of different small neighborhoods thus disappears, if we consider mono-functional protein families. At d>80 percent, however, neighborhood divergence becomes close to maximal, as in our earlier analysis. This is because protein pairs at this distance fall into different families, and typically have different functions. For example, a comparison of all pairs of monofunctional protein families within the TIM barrel domain shows that only 1.6 percent of these pairs have the same function. This pattern also holds for our whole data set, where 75 percent (1,162) of the protein families perform single functions and only 0.1 percent of the family pairs (with the same or different structure) have the same function.

In sum, if protein structure equaled function, then all but the most distant genotypic neighborhoods would be functionally homogeneous. Functional neighborhood diversity emerges from the multifunctionality of structures.

Discussion

In sum, our large data set of more than 30,000 protein sequences with known structures and enzymatic functions gives rise to three general observations. First, as shown previously [30], different functions are carried out by different numbers of sequences and structures. Second, most functions are restricted to single structures, but some can be carried out by many structures. Relatedly, most protein families are associated with only one function, as was also shown previously based on fewer data [30]. Third, and most important, different genotype neighborhoods tend to contain a different spectrum of functions, whose diversity increases with increasing distance of these neighborhoods in sequence space.

One would be more likely to find functions that can be executed by many structures in sequence space than those carried out by only one structure, because, with possible exceptions, such functions would also be carried out by more sequences. While it is tempting to interpret the first and second observation above as firm evidence that different functions differ in the proportion of sequences that can perform them, this evidence has to be taken with a grain of salt. First, some functions may be needed by few organisms or in few environments. Fewer proteins carrying out these functions may exist than for other, more generally important functions. Second, the data we analyze is not a random sample of sequence space. Some enzymes may be better studied than others, for reasons of their medical importance, or merely by historical accident. Fundamentally, every existing sample of proteins is subject to these problems. However, we can get hints about intrinsic differences among functions in the number of associated sequences if we study the number of functions per structure, in particular if we control for the different number of sequences per structure. Our analysis above showed that the number of structures per function has a nonuniform distribution, even after controlling for the number of known sequences for each structure (Figure S3). This observation hints that some functions may indeed be more frequent in sequence space than others.

In support of this notion, in vitro selection experiments on random polymers and mutagenesis experiments indeed suggest that proteins with different functions may occupy different proportions of sequence space [13,14,35]. For example, Taylor et al (2001) explored random libraries of a helical bundle chorismate mutase. They found previously unidentified residues involved in the formation of the enzyme active site. The authors estimate a probability of the order of 10^{-23} of finding this functional enzyme using the same fold in sequence space [13]. Axe [14] examined the probability to find an enzyme in sequence space. His results based on non-biased random libraries of beta-lactamase suggest that this catalyst is rare, with an occurrence probability of 10^{-64}. He suggests that the overall probability of finding any functional protein in the sequence space is as low as 10^{-77}. Yet another study used phage display to examine the probability to find ATP binding proteins from a random sample of sequence space regardless the fold [47]. Its authors estimated a probability of 10^{-11} to find an ATP binding protein, suggesting that a protein with this function could be found easily in a random search of the sequence space. Although estimates like these depend on various factors, including the length of the proteins considered, they suggest that the probability to find a functional protein in sequence space can vary broadly.

Our most important, third observation, the high phenotypic diversity of different neighborhoods in sequence space, has obvious implications for the evolution of novel protein functions. If a protein performs an essential function, then this function needs to be preserved over time. This typically means that the protein's structure will also be preserved, because changes in protein structure typically require changes in many amino acid sequences and would thus not preserve function [48,49]. Populations of organisms are subject to mutations that change individual amino acids. They may also be subject to recombination between homologous proteins of the closely related individuals within a population. This means that proteins that preserve their function change their genotype gradually over time. In other words, they drift through the function's genotype network, which can extend very far through genotype space [50,51]. In doing so, they explore different regions of genotype space, all the while preserving their function [52]. Consider now two proteins with the same function but in different parts of this space. If their neighborhoods typically contained the same spectrum of functions, the exploration of a genotype network would not aid in their exploration of novel functions. If conversely, these neighborhoods differ in the function they contain, the exploration of a genotype network may be crucial to explore new functions, some of which may become evolutionary innovations. This is exactly the property we found here. That is, by exploring a genotype network, proteins can explore ever-changing sequence neighborhoods, and an ever-changing spectrum of novel enzymatic functions.

The functional diversity of different neighborhoods we observe is caused by differences in the apparent structural promiscuity of a particular function. That is, if any one function could only be carried out by one structure, then different neighborhoods of two proteins with the same structure or function would not contain diverse novel functions. This observation underscores the importance of studying the organization of protein functions in sequence space independently from the organization of structures.

The phenotypic diversity of different neighborhoods in sequence space also has a flip side: It means that not all protein functions occur in every neighborhood of sequence space. In other words, the evolution of novel protein functions is *constrained* by an individual or a population's location in sequence space. A consequence of such constraints is evolutionary stasis, where genotypes but not phenotypes in a population change while the population explores a genotype network. Such stasis is interrupted by the discovery of novel phenotypes when a population arrives at a neighborhood where such novel phenotypes are found. In other words, evolutionary constraints can lead to patterns of episodic evolution, where periods of stasis are interrupted by discoveries of novel phenotypes. Such episodic evolution has been documented in systems ranging from evolving RNA molecules to macroscopic traits in the fossil record [53–57]. Although to our knowledge no demonstration of episodic evolution is known for protein functions, our observations suggest that it will also be widespread for proteins.

The causes of evolutionary constraints on the acquisition of new phenotypes are the subject of a broad literature and wide debate, particularly among students of organismal development and its evolution [58–62]. In this literature, the causes of constrained evolution are often unclear, because the relationship between genotype and phenotype is very complex for the macroscopic traits that development creates. This relationship involves many genes, and is thus incompletely understood. Protein functions are simpler, molecular phenotypes, which allow us to circumvent these complexities. For them, constrained evolution emerges from the organization of phenotypes in a genotype space. These observations, if generalizable to more complex traits, imply that we need to understand the organization of such complex traits in their genotype space, before we can hope to understand constrained evolution well.

Our study also reveals similarities and differences between the space of protein structure and functions when mapped onto sequence space (Figure 3, S2 and S13). As previous studies also showed, structures are highly conserved in sequence space [63,64]. For example, pairs of sequences may diverge by more than 95 percent and still fold into the same structure [11].

Early bioinformatic analyses suggested that the organization of protein functions was similar to that of protein structures [26–28], but later work showed that functions and structures have different organization in sequence space and functional annotation can not only rely on sequence similarity [32].

Here we observed that new functions are encountered at varying sequence distances as proteins diverge in sequence space, and that this property can be attributed to the fact that some

protein families perform multiple functions. While for short distances in sequences space this diversity is moderate, it increases at larger distances and once the structure conservation threshold (i.e. 70 to 80 percent sequence identity) is crossed, we observed an explosion in the accessibility of new structures [11,63], and consequently an enormous increase in functional diversity (Figure 3,4 and S13).

The characterization of protein sequence spaces with large but heterogeneous biological data like ours has several caveats. First, different proteins have different lengths, and thus exist in genotype spaces of different dimensions. To compare neighborhoods, however, we need to embed proteins within a genotype space of a given dimension. For our analysis, we solved this problem by restricting some analyses to proteins of similar length, and by focusing others on subsets of multiple sequence alignments that have the same lengths. This amounts to projecting genotype spaces of higher dimensions onto lower-dimensional spaces. It reduces the size of our data set, an unavoidable consequence of this procedure.

A second problem is posed by the vast size of genotype space. Our data set is very large, but even data sets many orders of magnitudes larger than ours would sample such a space only very sparsely. The limited functional diversity of the smallest sequence neighborhoods we examine likely results from this sparsity.

Third, our data set is a non-random sample of sequence space, with many biases whose extent is unknown. Some of the properties we study, such as the structural promiscuity of a function, are not easy to infer from such a data set, nor can they be inferred from models of protein folding such as lattice proteins, because such models are ill-suited to study protein function. We will not be able to characterize these properties rigorously until we are able to generate random samples in sequence space of proteins with a given function, which requires computational tools that are not yet within reach.

We note in closing that the property central to our study - the phenotypic diversity of different neighborhoods - is not likely to be strongly affected by biases in our data. Specifically, we showed that different phenotypic neighborhoods contain different phenotypes, largely because multifunctional protein structures exist. In our data, such multifunctional structures comprise a minority of structures. This observation may well be an artifact of a biased sampling of sequence space. If we had the same, large amount of sequence information for all structures, we might find most structures to be functionally versatile; and we might find most functions to be executable by multiple structures. If anything, the functional diversity of different neighborhoods in sequence space would thus increase. Thus, the very feature that both facilitates evolutionary exploration of novel functions and causes their constrained evolution is probably a generic property of protein sequence space.

Supporting Information

File S1 We extend earlier work on statistics of protein functions, specifically: 1) the number of structures per function for the six top-level EC functions; and 2) the numbers of sequences per function against the number of structures per function and the promiscuity of a function for the six major enzyme classes EC1 through EC6.
Found at: doi:10.1371/journal.pone.0014172.s001 (0.06 MB DOC)

Figure S1 Distribution of the number of sequences per structures and per functions. (a) Distribution of the number of sequences per structure. Histogram of the total number of sequences per structure (min = 1; max = 4.134; mean = 84). (b) Distribution of the number of sequences per function. Histogram of the total number of sequences per function, according to the EC classification finest-grained level (min = 1; max = 578; mean = 29). Distributions are based on our data set composed of 39,529 sequences, 457 structures and 1,343 enzymes types.
Found at: doi:10.1371/journal.pone.0014172.s002 (1.05 MB EPS)

Figure S2 Distribution of distances between sequences. (a) Distribution of distances between all sequence pairs with the same structure and function. (min = 0; max = 100; median = 55; mean = 54). The distribution shows values of all against all pairwise distances between sequences that fold into the same structure and are classified under the same enzyme function. (b) Distribution of distances between all sequence pairs with the same function. (min = 0; max = 100; median = 56; mean = 57). The functional annotation is based on the finest-grained level of the EC hierarchy. (c) Distribution of distances between all sequence pairs with the same structure. (min = 0; max = 100; median = 92; mean = 86). The data for these distributions was generated as follows. From our original data composed of 39,529 sequences, 457 structures and 1,343 enzyme functions, we extracted 10 independent samples of random sections from those multiple sequence alignments that comprised at least 100 amino acids. We required each random section to comprise 100 amino acids. These 10 samples were on average composed of 28,862 sequences, 337 structures and 1,036 enzyme functions. We then chose, from each of the 10 random samples, 10^7 sequence pairs with identical structure and/or function at random, and calculated their pairwise distances. Error bars indicate standard errors of the mean over the 10 independent samples.
Found at: doi:10.1371/journal.pone.0014172.s003 (0.99 MB EPS)

Figure S3 Distribution of the number of structures per function, corrected for the number of sequences. For this figure we used the original dataset of 39,529 sequences, 457 structures and 1,343 enzyme functions. We determined, for each structure i, the fraction f_i of sequences adopting this structure. For each function, we then determined all structures that are associated with this function, and averaged the corresponding values of f_i. The panel shows a histogram of these averages, for all 1,343 enzymatic functions.
Found at: doi:10.1371/journal.pone.0014172.s004 (0.01 MB EPS)

Figure S4 Structures per function versus sequences per function. Associations between number of sequences and structures per protein function at the fourth, finest-grained (a,b) and the first, coarsest level (c,d) of the EC hierarchy. For the first analysis (panel a and b), we classified the 39,529 sequences of our original data set according to their enzyme functions and compared the number of sequences per function with the number of structures per function. There are a total of 457 structure and 1,343 functions at this level. For the second analysis of the top-level EC functions, the 39,529 sequences fall into only 6 different enzyme types. While it is difficult to make statistically rigorous statements based on so few functions, we nonetheless wanted to understand how sensitive our observations in panel c) and d) were to the structure of our data. To this end, we extracted random samples of 10^4 sequences from our data set and classified them according to the 6 top EC-levels. We repeated this procedure 10^5 times and compare the statistics of the averaged values obtained from the sampling with the statistics observed for the whole data set (without sampling). Plots show the means over the sampling and error bars the standard deviations. (a) Scatterplot of the number of sequences per function against the number of structures per function. Spearman rank's correlation r = 0.29 (P<E-50). (b) Scatterplot of the number of sequences per function versus structural promiscuity. Spearman rank's correla-

tion r = 0.27 (P<E-50). (c) Scatterplot of the number of sequences per function against the number of structures per function at the top level of the EC hierarchy. Spearman rank's correlation r = 0.92 (P<0.01). Spearman rank's correlation of the complete data set (without sampling) is r = 0.94 (P<0.01). (d) Scatterplots of the number of sequences per function at the coarsest level of the EC hierarchy versus structural promiscuity. Spearman rank's correlation r = 0.92 (P<0.01). Note the decadic logarithms on the vertical axes of all plots. Spearman rank's correlation of the complete data set (without sampling) is r = 0.77 (P<0.1).
Found at: doi:10.1371/journal.pone.0014172.s005 (1.70 MB DOC)

Figure S5 Distribution of structures over functions at the top level of the EC hierarchy. (a) Number of structures per enzyme class at the first (top) level of the EC hierarchy. For this figure, we grouped the total number of different structures (457) in our dataset composed of 39,529 sequences are classified according to the enzyme function that they perform (min = 28; max = 188; mean = 100). (b) Structural promiscuity at the first level of the EC hierarchy. Structural promiscuity (R_F) is an entropy-like measure (see main text of the Supplementary Material) calculated from the distribution of the EC top-level types of enzyme functions over different protein structures (min = 0.32; max = 0.57; mean = 0.49).
Found at: doi:10.1371/journal.pone.0014172.s006 (0.88 MB EPS)

Figure S6 Distribution of functions over structures.(a) Distribution of the number of functions per structure at the fourth (finest grained) level of the EC hierarchy. (min = 1, max = 103). (b) Distribution of functional versatility (V_S) at the fourth level of the EC hierarchy. Functional versatility (V_S) is an entropy-like measure (see main text) calculated from the distribution of structure domains over different enzyme functions at the bottom level of the EC hierarchy. (min = 0, max = 0.53). For the data in these panels, we classified the total number of different enzyme functions (1,343) according to the structures that carry them out (457).
Found at: doi:10.1371/journal.pone.0014172.s007 (1.04 MB EPS)

Figure S7 Distribution of functions over structures at the coarsest level of the EC hierarchy.(a) Distribution of the number of functions per structure at the coarsest level of the EC hierarchy. The data is based on the total number of 6 different enzyme types at the first, coarsest level of the EC hierarchy in our dataset of 39,529 sequences and 457 strcutures. For the plot, we classified each sequence according to its structure and function. (min = 1, max = 5;). (b) Distribution of functional versatility (V_S) at the coarsest level of the EC hierarchy. Functional versatility (V_S) is an entropy-like measure (see text) calculated here from the distribution of structure domains over different enzyme functions at the first, coarsest level of the EC hierarchy (min = 0, max = 0.76). The inset show the same data, but with a $\log_{10}$-transformed vertical axis.
Found at: doi:10.1371/journal.pone.0014172.s008 (0.86 MB EPS)

Figure S8 Sequences per structure versus the distribution of functions. (a) Scatterplot of the number of sequences per structure against the number of functions per structure. The association between number of sequences and enzyme functions per structure domain is shown for the fourth (finest grained) level of the EC hierarchy. Spearman rank's correlation r = 0.57 (P<E-50). (b) Scatterplot of the number of sequences per structure versus functional versatility. The same dataset described in panel (a) is used to examine the association between number of sequences (39,529) and the functional versatility (V_S) per structure domain. Spearman rank's correlation r = 0.51 (P<E-50). For the data in this figure, we classified the number of sequences (39,529) and enzyme functions (1,343) according to their structure (457). Note the log10-transformed horizontal axes.
Found at: doi:10.1371/journal.pone.0014172.s009 (1.33 MB EPS)

Figure S9 Scatterplot of the number of sequences per structure. Associations between numbers of sequences and functions per structure are shown at the first, coarsest level of the EC hierarchy. We classified the 39,529 sequences according to their 457 structures and compared the number of sequences per structure with (a) the number of functions per structure and (b) functional versatility (V_S). For the first analysis (panel a), we classified the number of functions (at the coarsest level of the EC hierarchy) per structure in our dataset and the corresponding number of sequences folding into those structures (Spearman rank's correlation r = 0.43; P<E-50), Error bars represent the standard error over the number of sequences per structure. The second panel (b) shows a scatterplot comparing the number of sequences per structure ($\log_{10}$-transformed) and V_S per structure (Spearman rank's correlation r = 0.42; P<E-50).
Found at: doi:10.1371/journal.pone.0014172.s010 (0.98 MB EPS)

Figure S10 Principal Component Analysis (PCA) of the TIM barrel main homologous superfamily (the aldolase I superfamily). For this analysis, we first constructed a multiple sequence alignment of the aldolase I superfamily (CATH code: 3.20.20.70), using the program clustalw, and allowing no more than 10 percent gaps in the alignment. The resulting multiple sequence alignment is composed of 4,132 sequences of length 188 amino acids, and comprises 53 different enzyme functions at the finest-grained level of the EC hierarchy. For subsequent PCA [4], we encoded the sequences in the alignment as numeric strings (21 possible values per amino acid position, including gaps). The panels show the first two principal components (a) and the first and third components (b). The 53 different enzyme functions are color-coded according to the color bar to the right. Note the clear separation of some functions.
Found at: doi:10.1371/journal.pone.0014172.s011 (3.82 MB EPS)

Figure S11 Genotypic neighborhoods of proteins with a given structure. The figure shows the dependency between the radius and distance of sequence neighborhoods, and the fraction F_u of functions unique to one neighborhood, for sequences folding into 36 different structures. The total set of multiple alignments we used in this analysis comprises a total of 18,117 sequences with lengths ranging from 100 to 400 amino acids, and spans 434 enzymatic functions covering all 6 EC classes. We analysed these sequences exhaustively. That is, for all possible pairwise sequence comparisons we computed their values of r, d and F_u. The heatmap shows F_u values at each combination of d and r, for the 26 structures (a) Heatmap of the fraction of unique functions (F_u) at different combinations of neighborhood radii (r) and sequences distances (d). (b) Fraction of unique functional F_u of unique functions versus sequence distance (expressed in percent) at a given neighborhood radius, as shown in the legend. Due to the sparsity of data, we grouped values into 20 different distance bins, each spanning d = 5. Error bars represent standard errors calculated for these 20 bins. The CATH identifiers of the 36 superfamilies we used in this analysis are listed here: 3.30.70.141; 3.30.420.10; 3.40.50.960; 2.70.40.10; 3.90.45.10; 3.40.50.2020; 3.20.19.10; 3.40.50.1470; 3.40.50.1360; 2.40.10.10; 3.90.1550.10; 3.90.226.10; 3.90.180.10; 3.40.50.880; 3.60.20.10; 3.40.50.620; 3.40.1210.10; 3.40.1160.10; 3.40.50.1240; 3.40.640.10; 3.60.15.10; 3.20.20.60; 3.20.20.70;

3.30.572.10; 3.90.550.10; 1.20.200.10; 3.40.1190.20; 3.30.930.10; 1.10.1040.10; 3.20.20.140; 3.40.50.1820; 3.20.20.210; 3.20.20.150; 3.40.718.10; 3.20.20.80; 1.10.630.10.
Found at: doi:10.1371/journal.pone.0014172.s012 (2.31 MB EPS)

Figure S12 Distribution of the number of protein families per structures. (a) Distribution of the number of protein families per structure domain in the whole CATH database. This data is composed of 114,215 protein families grouped into 2,178 structures. (b) Distribution of the number of protein families per structure in our dataset composed of 39,529 sequences and 457 structures. More precisely, the notion of a protein family here corresponds to that of a CATH homologous superfamily (Greene et al, 2007). The insets show the same data, but with a $\log_{10}$-transformed vertical axis.
Found at: doi:10.1371/journal.pone.0014172.s013 (0.92 MB EPS)

Figure S13 Neighborhood diversity in functions depends on functionally versatile protein families and structures. The figure shows the dependency between the radius and distance of two genotype neighborhoods, and the fraction F_u of functions unique to one neighborhood. (a) Heatmap of the fraction of unique functions (F_u) at different combinations of neighborhood radii (r) and sequences distances (d). The data is based on the major superfamily of the TIM barrel domain, aldolase I (CATH code: 3.20.20.70), which is composed of 4,132 sequences that carry out 53 different enzyme functions (see methods). These sequences can be grouped into 62 protein families. From this data set we selected the 30 protein families that carry out single enzyme functions. These families comprise 2,444 protein sequences and 27 enzyme functions. For all possible sequence pairs in this data set we computed values of d and F_u for different values of r. The heatmap shows Fu values over all distance-radius combinations. (b) Fraction of unique functional variations versus sequence distance (expressed in percent) at constant neighborhood radii, as shown in the legend. Due to the sparsity of the data, we grouped values into 20 different distance bins, each spanning $d = 5$. Error bars represent standard errors calculated for these 20 bins.
Found at: doi:10.1371/journal.pone.0014172.s014 (2.09 MB EPS)

Acknowledgments

EF thanks Margot Crucet for insightful discussions.

Author Contributions

Conceived and designed the experiments: EF AW. Performed the experiments: EF. Analyzed the data: EF. Wrote the paper: EF AW.

References

1. Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225: 563–564.
2. Hamming RW (1980) Coding and Information Theory, Prentice Hall, Englewook Cliffs, N.J.
3. Mantaci S, Restivo A, Sciortino M (2008) Distance measures for biological sequences: Some recent approaches. Int J Approximate Reasoning 47: 109–124.
4. Finkelstein AV, Gutin AM, Badretdinov AY (1995) Boltzmann-like statistics of protein architectures. Origins and consequences. In: Biswas BB, Roy S, eds. Subcellular Biochemistry, Vol 24. Proteins: Structure, function and engineering. Plenum Press, New York.
5. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372: 631–634.
6. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
7. Kleina L, Miller J (1990) Genetics studies of the lac repressor. 13. Extensive amino-acid replacements generated by the use of natural and synthetic nonsense suppressors. J Mol Biol 212: 295–318.
8. Rennell D, Bouvier S, Hardy L, Poteete A (1991) Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 222: 67–87.
9. Huang W, Petrosino J, Hirsch M, Shenkin P, Palzkill T (1996) Amino acid sequence determinants of beta-lactamase structure and activity. J Mol Biol 258: 688–703.
10. Aronson HE, Royer WE, Jr., Hendrickson WA (1994) Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci 3: 1706–1711.
11. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85–94.
12. Kunin V, Teichmann SA, Huynen MA, Ouzounis CA (2005) The properties of protein family space depend on experimental design. Bioinformatics 21: 2618–2622.
13. Taylor SV, Walter KU, Kast P, Hilvert D (2001) Searching sequence space for protein catalysts. Proc Nat Acad Sci USA 98: 1056–1060.
14. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. J Mol Biol 341: 1295–1315.
15. Li H, Helling R, Tang Ch, Wingreen N (1996) Emergence of preferred structures in a simple model of protein folding. Science 273: 666–669.
16. Kussell E (2005) The designability hypothesis and protein evolution. Protein Pept Lett 12: 111–116.
17. Brannigan JA, Wilkinson AJ (2002) Protein engineering 20 years on. Nat Rev Mol Cell Biol 3: 964–970.
18. Michael SF, Kilfoil VJ, Schmidt MH, Amann BT, Berg JM (1992) Metal binding and folding properties of a minimalist Cys2His2 Zinc finger peptide. Proc Natl Acad Sci USA 89: 4796–4800.
19. Choo Y, Isalan M (2000) Advances in zinc finger engineering. Current Opinion in Structural Biology 10: 411–416.
20. Buchler NEG, Goldstein RA (1999) Effect of alphabet size and foldability requirements on protein structure designability. Proteins Struct Funct Bioinf 34: 113–124.
21. Mann M, Backofen R, Will S (2009) Equivalence classes of optimal structures in HP protein models including side chains. In Proceedings of the Fifth Workshop on Constraint Based Methods for Bioinformatics (WCB09).
22. Bornberg-Bauer E (1997) How are model protein structures distributed in sequence space? Biophys J 73: 2393–2403.
23. Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci USA 96: 10689–10694.
24. Holm L, Sander C (1998) Touring protein fold space with Dali/FSSP. Nucleic Acids Res 26: 316–319.
25. Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. Curr Opin Struct Biol 14: 202–207.
26. Shah I, Hunter L (1997) Predicting enzyme function from sequence: a systematic appraisal. In Fifth International Conference on Intelligent Systems for Molecular Biology. In: Gaasterland T, Karp P, Karplus K, Ouzounis C, Sander C, Valenvia A, eds. 276–283, Halkidiki, Greece: AAAI Press.
27. Pawlowski K, Jaroszewski L, Rychlewski L, Godzik A (2000) Sensitive sequence comparison as protein function predictor. Pac Symp Biocomp 8: 42–53.
28. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.
29. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297: 233–249.
30. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143.
31. Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 8: 294.
32. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.
33. Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321: 741–765.
34. Steitz TA (1999) DNA polymerases: Structural diversity and common mechanisms. J Biol Chem 274: 17395–19398.
35. Reidhaar-Olson JF, Sauer RT (1990) Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. Proteins 7: 306–316.
36. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shape and back: a case study in RNA secondary structures. Proc R Soc London, Ser B 255: 279–284.
37. Ciliberti S, Martin OC, Wagner A (2007) Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comp Biol 3(2): e15.
38. Matias Rodrigues JF, Wagner A (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comp Biol 5(12): e1000613.
39. The UniProt Consortium (2009) The Universal Protein Resource (UniProt). Nucleic Acids Res 37: D169–D174.
40. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–D297.

41. Eddy SR (1998) Profile Hidden Markov Models. Bioinformatics 14: 755–763.
42. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28: 304–305.
43. Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
44. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–17.
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
46. Prakash S, Johnson RE, Prakash L (2005) Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. Annu Rev Biochem 74: 317–353.
47. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410: 715–718.
48. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, et al. (1998) Protein folds and functions. Structure 6: 875–884.
49. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 288: 147–164.
50. Lipman DJ, Wilbur WJ (1991) Modelling neutral and selective evolution of protein folding. Proc R Soc London, Ser B 245: 7–11.
51. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, et al. (2001) Exploring Protein Sequence Space Using Knowledge-based Potentials. J Theor Biol 212: 35–46.
52. Aharoni A, Gaidukov L, Khersonsky O, Gould McQ S, Roodveldt C, et al. (2005) The evolvability of promiscuous protein functions. Nature Genetics 37: 73–76.
53. Knoll AH (1992) The early evolution of eukaryotes - a geological perspective. Science 256: 622–627.
54. Elena SF, Cooper VS, Lenski RE (1996) Punctuated evolution caused by selection of rare beneficial mutations. Science 272: 1802–1804.
55. Fontana W, Schuster P (1998a) Continuity in evolution: On the nature of transitions. Science 280: 1451–1455.
56. Fontana W, Schuster P (1998b) Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol 194: 491–515.
57. Adams KL, Qiu YL, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Nat Acad Sci USA 99: 9905–9912.
58. Stenseth NC, Maynard Smith J (1984) Coevolution in ecosystems: red queen evolution or stasis. Evolution 38: 870–880.
59. Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, et al. (1985) Developmental constraints and evolution. Q Rev Biol 60: 265–287.
60. Amundson R (1994) Two Concepts of constraint - adaptationism and the challenge from developmental biology. Philosophy of Science 61: 556–578.
61. Hodin J (2000) Plasticity and constraints in development and evolution. Presented at Modularity of Animal Form Workshop, Friday Harbor, Washington.
62. Brakefield PM (2006) Evo-devo and constraints on selection. Trends Ecol Evol 21: 362–368.
63. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823–6.
64. Chothia C, Lesk AM (1987) The evolution of protein structures. Cold Spring Harb Symp Quant Biol 52: 399–405.

4.1 Supplementary Material

Supplementary Results

We here analyze the number of structures per function for the six top-level EC functions given in Figure S5a. It can be seen that certain enzyme functions are more common than others. This bias may stem from biases in experimental studies, such as differences in the ease with which soluble proteins crystallize. However, some authors have argued that certain enzyme chemistries would prevail over others due to their requirements in nature (Nagano et al. 2001). We also characterized functional promiscuity for the six major enzyme classes (Figure S5b). Its distribution is nearly even among classes, with a small excess for hydrolases (EC.3). These are enzymes that cleave molecular bonds using water, and include proteases and lipases. Hydrolases are the most common enzyme class in the EC commission nomenclature, and also have more subclasses than other classes. We note that high versatility of a function on one level does not imply high versatility on another level. For example, the most abundant DNA polymerase function we discussed earlier is not a member of the most abundant enzyme class of hydrolases (EC.3). Instead, it is a transferase.

Figure S4c and d plot the numbers of sequences per function against the number of structures per function (panel c), and the promiscuity of a function (panel d) for the six major enzyme classes EC1 through EC6. For only six classes, it is difficult to calculate statistically meaningful associations between the numbers of sequences per function, and measures of functional promiscuity.

Most structures carry out few enzymatic functions.

Our earlier analyses (Figures S1-S5) focused on individual functions, and asked by how many structures they are carried out. In a complementary analysis, we now focus on individual structures, and ask about the number of functions they carry out. This analysis extends earlier work on statistics of protein functions (Martin et al. 1998; Nagano et al. 2001; Todd et al. 2001). The result is shown as a histogram for our 1,343 lowest-level enzyme functions in Figure S6a. It shows that most structures (54 percent; 248 structures) carry out only one function. The structure with the largest number of 103 associated functions is the NAD(P) binding Rossmann-like domain.

A complementary measure of the extent to which the same structure carries out different functions considers only sequences that have a given structure. It denotes by *f(i)* the fraction of these sequences with function *i*, and defines a normalized measure of how functionally versatile a structure is. We call this measure the *functional versatility* (V_S) of a structure, and define it, exactly as our above measure of a function's promiscuity, as $V_S = [-\sum_{i=1, f(i)\neq 0}^{i=N} f(i) \ln f(i)] / \ln N$. This measure is again akin to a normalized entropy. It ranges from a value of zero if all sequences with this structure carry out a single function, to a maximal value of one if *f(i)* has the same value for all *i* functions, that is, if a randomly chosen sequence with this structure were equally likely to adopt any function. Figure S6b shows the distribution of this promiscuity measure for our 457 structures. This distribution is again is highly skewed. It has its lowest value of 0 for 248 structures associated with only a single function, as well as a maximum value of 0.53, which occurs for the TIM barrel domain.

Figure S7 shows analogous analyses, but only for the 6 coarsest levels of the EC hierarchy. A major proportion of the structures (79 percent) carry out functions that fall into a single functional class. Only two of the 457 structures are associated with functions in 5 of the 6 major enzyme classes. These are the Rossmann and the TIM barrel homologous superfamilies. Figure S7b shows the distribution of functional versatility V_S for the six top-level enzyme functions. As expected, this distribution is highly skewed.

As in our other analyses, we wanted to explore the extent to which these distributions result simply from biases in available amounts of sequence information. Figure S8 shows, for 1,343 enzymatic functions, scatterplots of the number of sequences per structure against the number of functions per structure (panel a), and against the versatility of a structure (panel b). The figure shows that both the number of functions per structure, and the apparent versatility of a structure increase with the number of sequences that are associated with a function. These observations again suggest that apparent low versatility of a structure may result from a limited number of characterized sequences with this structure. Figure S9a and b show analogous scatterplots for the six top-level enzymatic functions. A positive association is also evident here.

References

Jolliffe IT. 2002. Principal component analysis. Springer Verlag, New York, Inc.

Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA. 1998. Protein folds and functions. Structure 6:875-884.

Nagano N, Porter CT, Thornton JM. 2001. The (beta-alpha)8 glycosidases: sequence and structure analyses suggest distant evolutionary relationships. Protein Eng. 14:845-855.

Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307:1113-1143.

Supplementary Figures

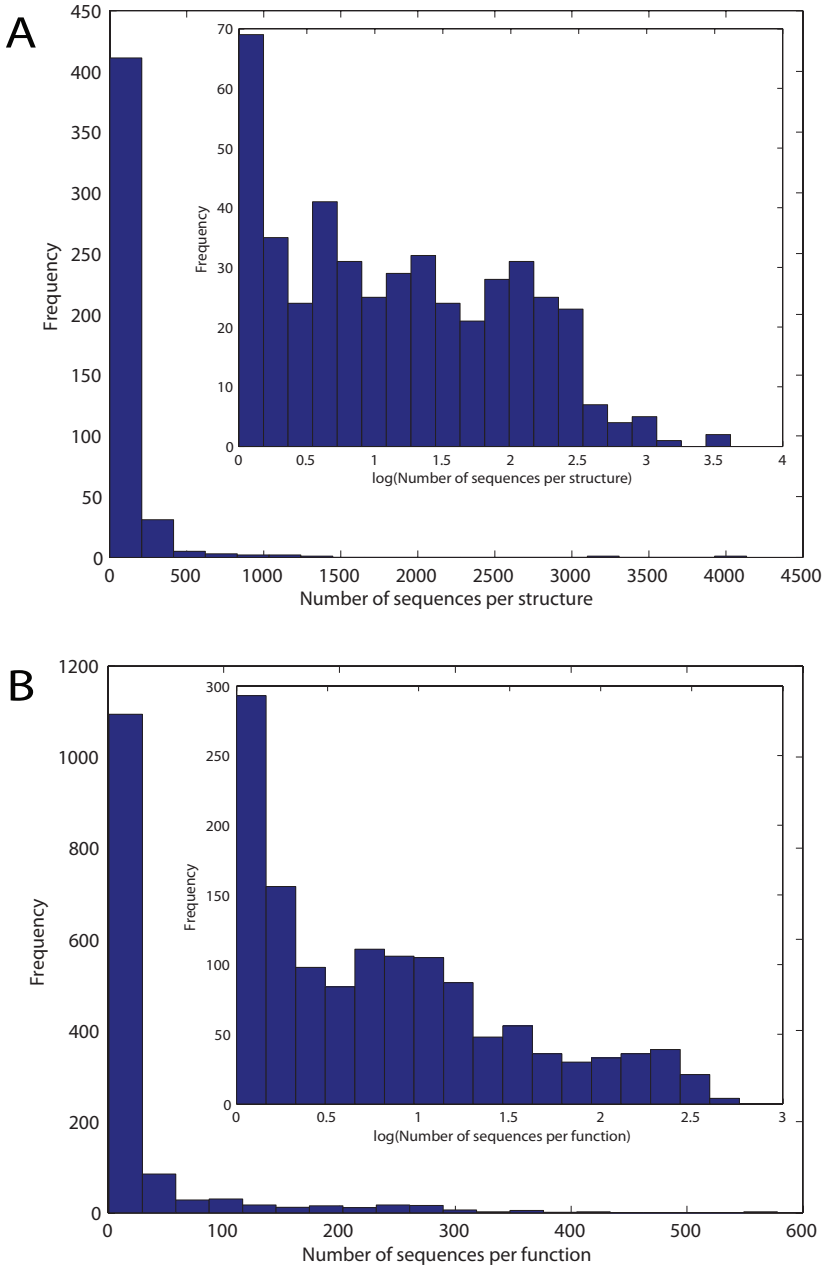
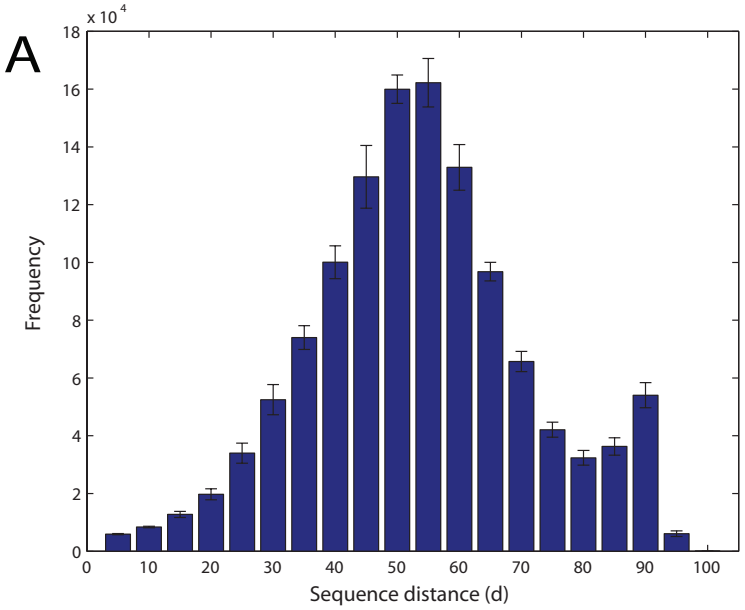
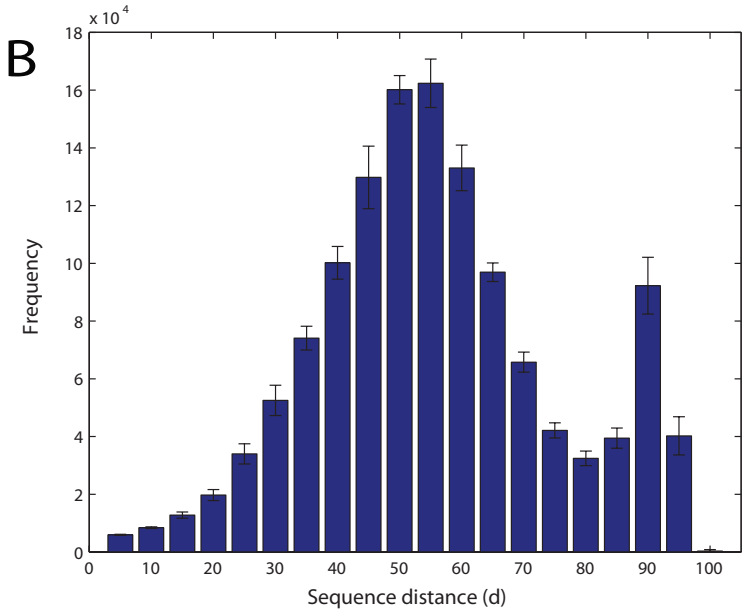


Figure S1. (**a**) *Distribution of the number of sequences per structure*. Histogram of the total number of sequences per structure (min=1; max=4.134; mean=84). (**b**) *Distribution of the number of sequences per function*. Histogram of the total number of sequences per function, according to the EC classification finest-grained level (min=1; max=578; mean=29). Distributions are based on our data set composed of 39,529 sequences, 457 structures and 1,343 enzymes types.


A
x 10^4
Frequency
Sequence distance (d)


B
x 10^4
Frequency
Sequence distance (d)

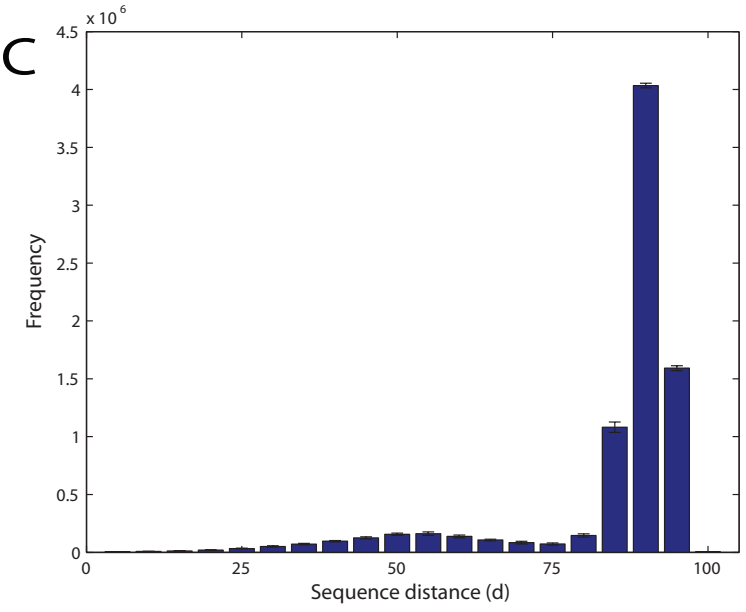

C
x 10^6
Frequency
Sequence distance (d)

Figure S2. *Distribution of distances between sequences*. (**a**) Distribution of distances between all sequence pairs with the same structure and function. (min=0; max=100; median=55; mean=54). The distribution shows values of all against all pairwise distances between sequences that fold into the same structure and are classified under the same enzyme function. (**b**) Distribution of distances between all sequence pairs with the same function. (min=0; max=100; median=56; mean=57). The functional annotation is based on the finest-grained level of the EC hierarchy. (**c**) Distribution of distances between all sequence pairs with the same structure. (min=0; max=100; median=92; mean=86). The data for these distributions was generated as follows. From our original data composed of 39,529 sequences, 457 structures and 1,343 enzyme functions, we extracted 10 independent samples of random sections from those multiple sequence alignments that comprised at least 100 amino acids. We required each random section to comprise 100 amino acids. These 10 samples were on average composed of 28,862 sequences, 337 structures and 1,036 enzyme functions. We then chose, from each of the 10 random samples, 10^7 sequence pairs with identical structure and/or function at random, and calculated their pairwise distances. Error bars indicate standard errors of the mean over the 10 independent samples.

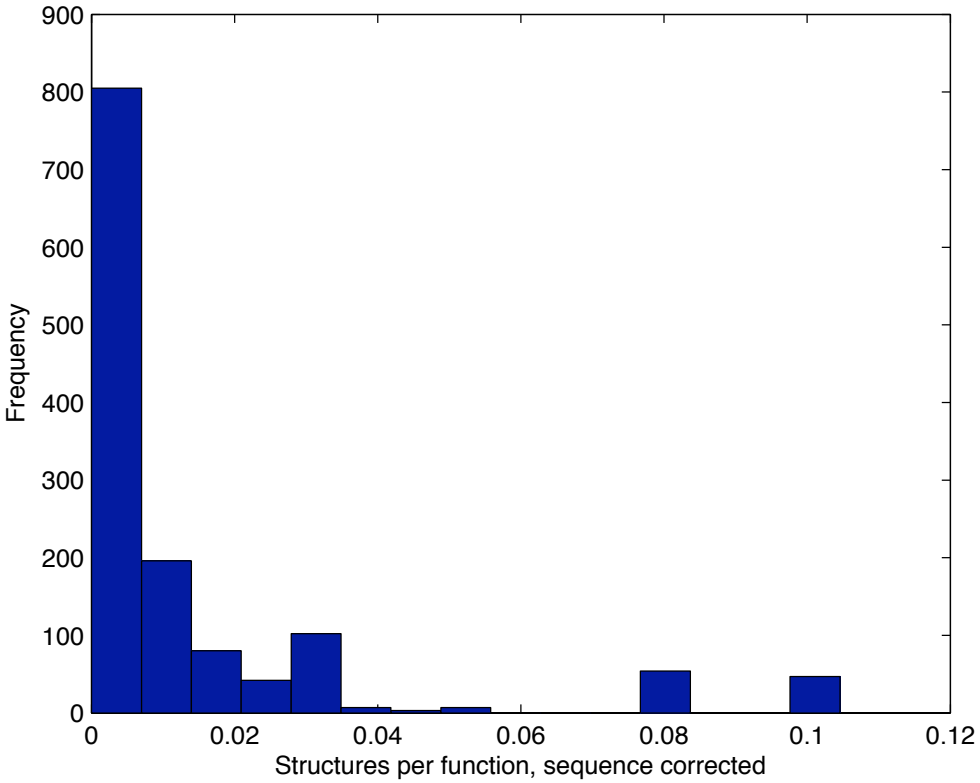


Figure S3. *Distribution of the number of structures per function, corrected for the number of sequences.* For this figure we used the original dataset of 39.529 sequences, 457 structures and 1,343 enzyme functions. We determined, for each structure i, the fraction f_i of sequences adopting this structure. For each function, we then determined all structures that are associated with this function, and averaged the corresponding values of f_i. The panel shows a histogram of these averages, for all 1,343 enzymatic functions.

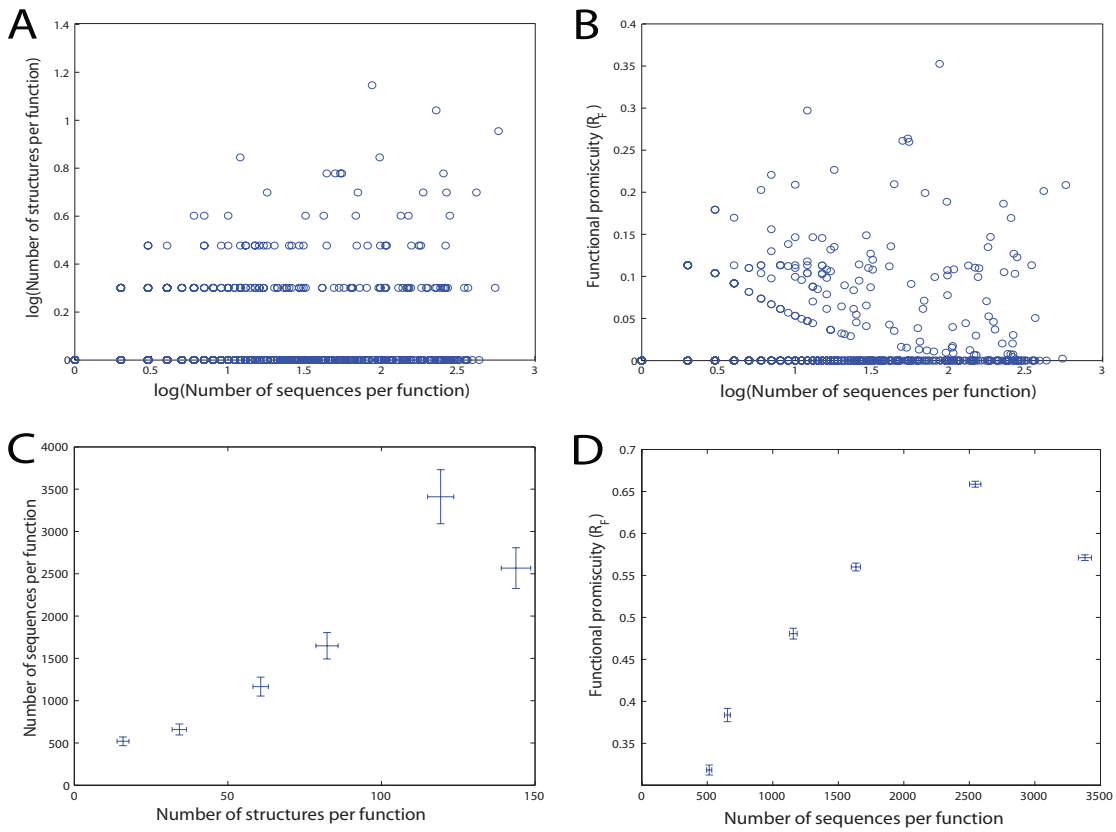


Figure S4. *Structures per function versus sequences per function.* Associations between number of sequences and structures per protein function at the fourth, finest-grained (*a,b*) and the first, coarsest level (*c,d*) of the EC hierarchy. For the first analysis (panel a and b), we classified the 39,529 sequences of our original data set according to their enzyme functions and compared the number of sequences per function with the number of structures per function. There are a total of 457 structure and 1,343 functions at this level. For the second analysis of the top-level EC functions, the 39,529 sequences fall into only 6 different enzyme types. While it is difficult to make statistically rigorous statements based on so few functions, we nonetheless wanted to understand how sensitive our observations in panel c) and d) were to the structure of our data. To this end, we extracted random samples of 10,000 sequences from our data set and classified them according to the 6 top EC- levels. We repeated this procedure 10^5 times and compare the statistics of the averaged values obtained from the sampling with the statistics observed for the whole data set (without sampling). Plots show the means over the sampling and error bars the standard deviations. (**a**) *Scatterplot of the number of sequences per function against the number of structures per function*. Spearman rank's correlation r=0.29 (P<E-50). (**b**) *Scatterplot of the number of sequences per function versus function promiscuity.*

Spearman rank's correlation r=0.27 (P<E-50). (**c**) *Scatterplot of the number of sequences per function against the number of structures per function at the top level of the EC hierarchy*. Spearman rank's correlation r=0.92 (P<0.01). Spearman rank's correlation of the complete data set (without sampling) is r=0.94 (P<0.01). (**d**) *Scatterplots of the number of sequences per function at the coarsest level of the EC hierarchy versus function promiscuity.* Spearman rank's correlation r=0.92 (P<0.01). Note the decadic logarithms on the vertical axes of all plots. Spearman rank's correlation of the complete data set (without sampling) is r=0.77 (P<0.1).

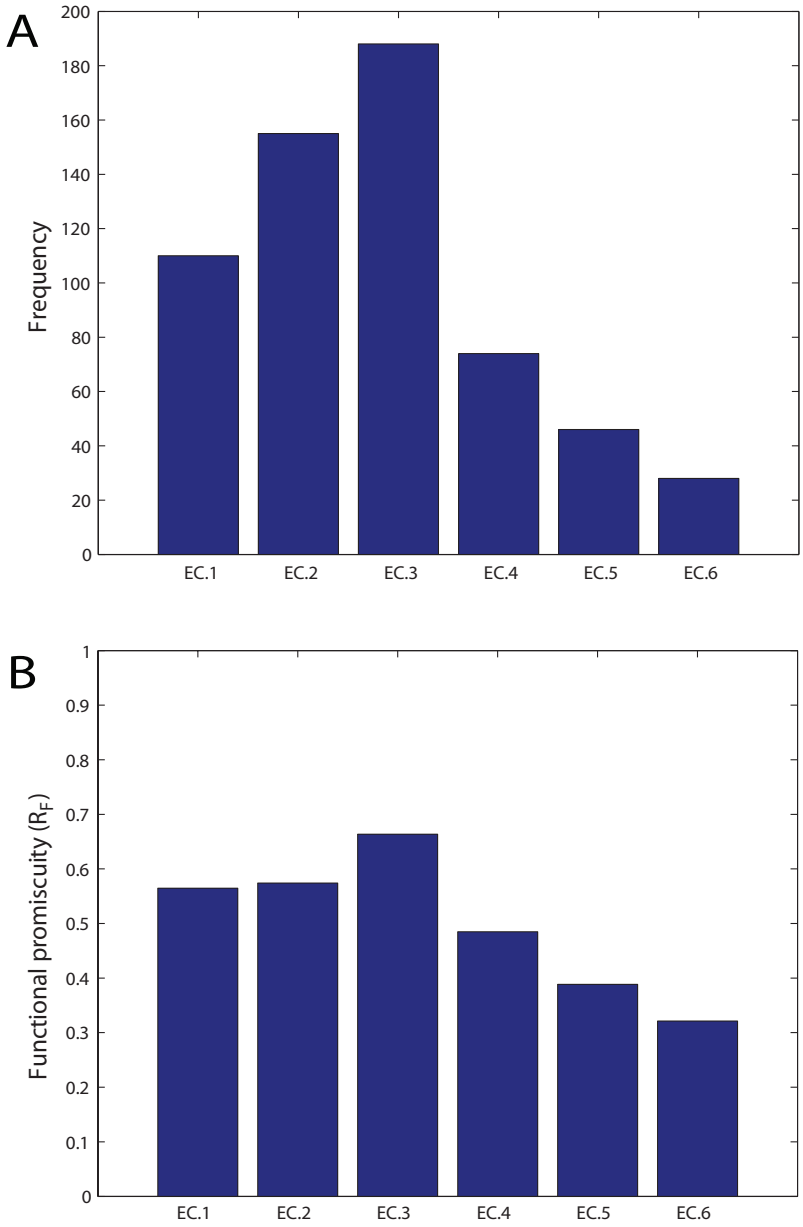


Figure S5. (**a**) *Number of structures per enzyme class at the first (top) level of the EC hierarchy*. For this figure, we grouped the total number of different structures (457) in our dataset composed of 39.529 sequences are classified according to the enzyme function that they perform (min=28; max=188; mean= 100). (**b**) Functional *promiscuity at the first level of the EC hierarchy*. Functional promiscuity (R_F) is an entropy-like measure (see main text of the Supplementary Material) calculated from the distribution of the EC top-level types of enzyme functions over different protein structures (min=0.32; max=0.57; mean=0.49).

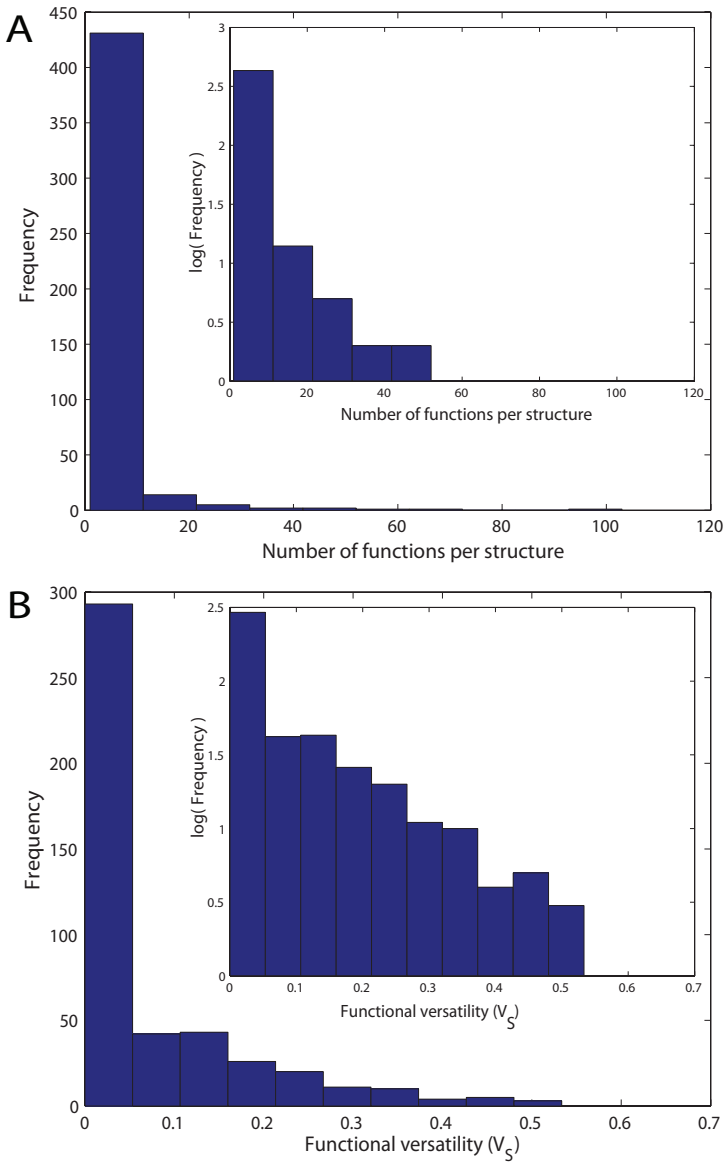


Figure S6. (**a**) *Distribution of the number of functions per structure at the fourth (finest grained) level of the EC hierarchy.* (min=1, max=103). (**b**) *Distribution of functional versatility (V_S) at the fourth level of the EC hierarchy.* Functional versatility (V_S) is an entropy-like measure (see main text) calculated from the distribution of structure domains over different enzyme functions at the bottom level of the EC hierarchy. (min=0, max=0.53). For the data in these panels, we classified the total number of different enzyme functions (1,343) according to the structures that carry them out (457).

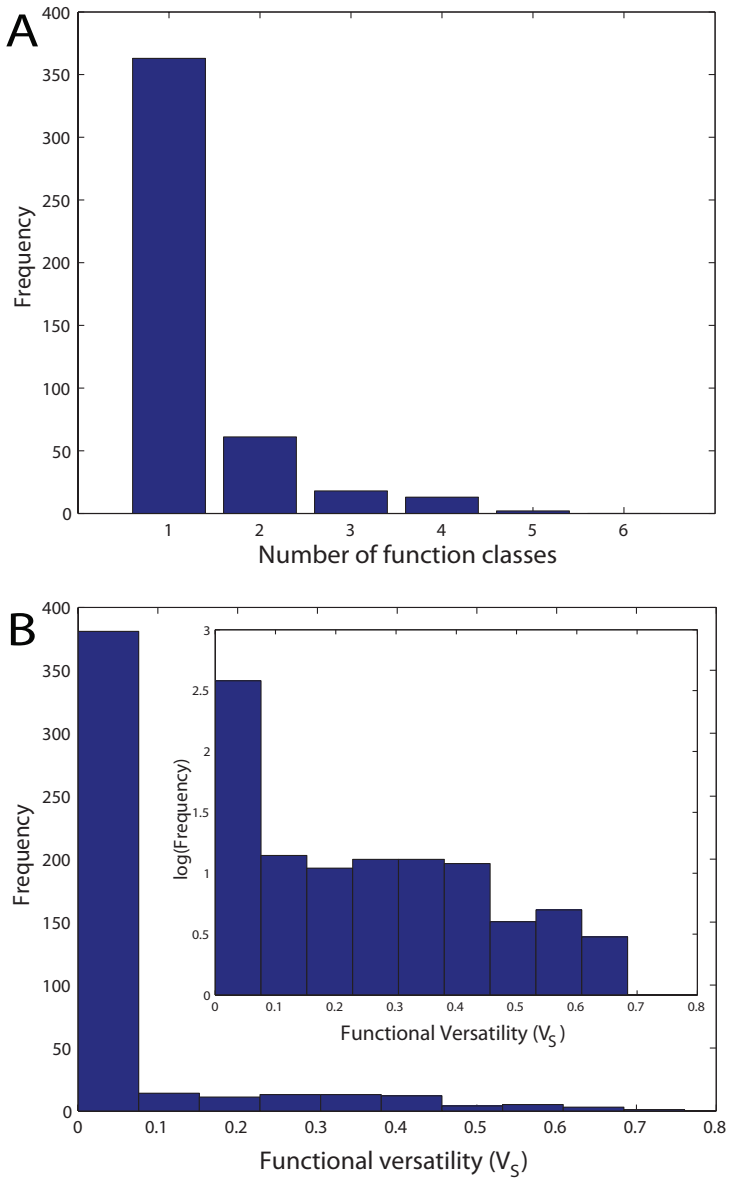


Figure S7. (**a**) *Distribution of the number of functions per structure at the coarsest level of the EC hierarchy.* The data is based on the total number of 6 different enzyme types at the first, coarsest level of the EC hierarchy in our dataset of 39.529 sequences and 457 strcutures. For the plot, we classified each sequence according to its structure and function. (min=1, max=5;). (**b**) *Distribution of functional versatility (V_S) at the coarsest level of the EC hierarchy.* Functional versatility (V_S) is an entropy-like measure (see text) calculated here from the distribution of structure domains over different enzyme functions at the first, coarsest level of the EC hierarchy. (min=0, max=0.76). The inset show the same data, but with a $\log_{10}$-transformed vertical axis.

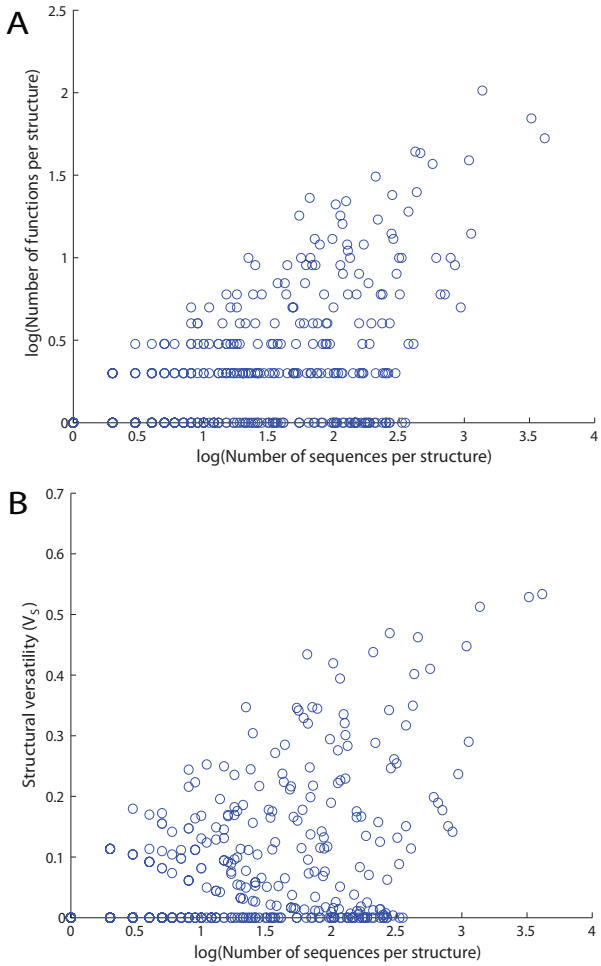


Figure S8. (**a**) *Scatterplot of the number of sequences per structure against the number of functions per structure*. The association between number of sequences and enzyme functions per structure domain is shown for the fourth (finest grained) level of the EC hierarchy. Spearman rank's correlation r=0.57 (P<E-50). (**b**) *Scatterplot of the number of sequences per structure versus functional versatility.* The same dataset described in panel (a) is used to examine the association between number of sequences (39,529) and the functional versatility (V_S) per structure domain. Spearman rank's correlation r=0.51 (P<E-50). For the data in this figure, we classified the number of sequences (39,529) and enzyme functions (1,343) according to their structure (457). Note the $\log_{10}$-transformed horizontal axes.

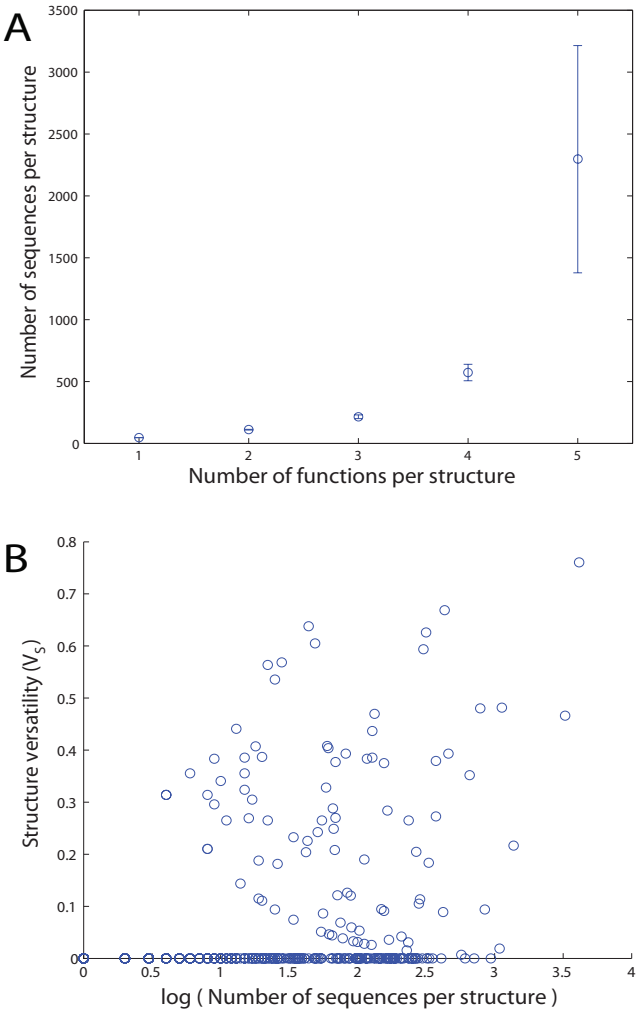


Figure S9. *Scatterplot of the number of sequences per structure.* Associations between numbers of sequences and functions per structure are shown at the first, coarsest level of the EC hierarchy. We classified the 39,529 sequences according to their 457 structures and compared the number of sequences per structure with (a) the number of functions per structure and (b) structural versatility (V_S). For the first analysis (panel a), we classified the number of functions (at the coarsest level of the EC hierarchy) per structure in our dataset and the corresponding number of sequences folding into those structures (Spearman rank's correlation r=0.43; P<E-50), Error bars represent the standard error over the number of sequences per structure. The second panel (b) shows a scatterplot comparing the number of sequences per structure (log_{10}-transformed) and V_S per structure (Spearman rank's correlation r= 0.42; P<E-50).

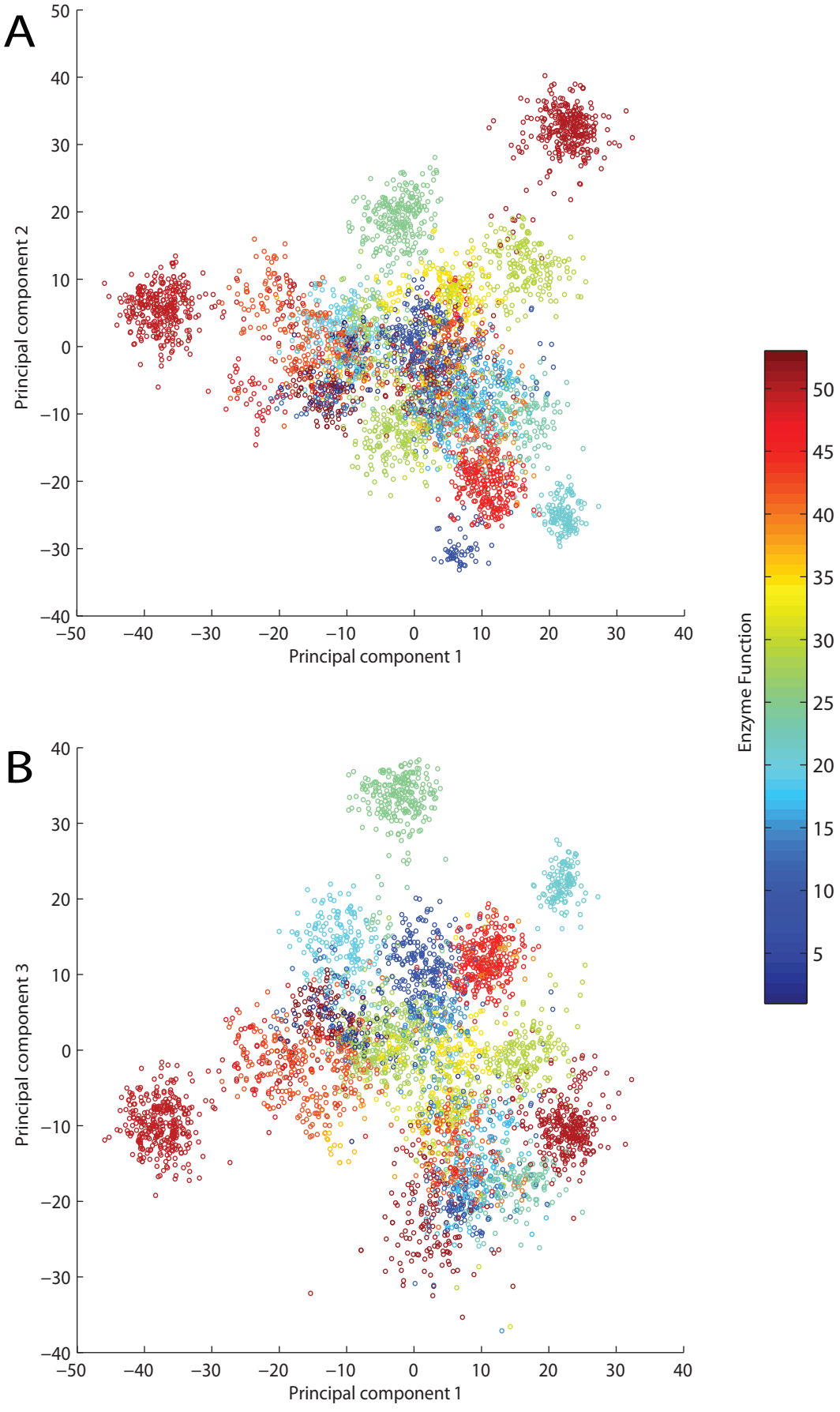

A
Principal component 2
Principal component 1
B
Principal component 3
Principal component 1
Enzyme Function

Figure S10. *Principal Component Analysis (PCA) of the TIM barrel main homologous superfamily (the aldolase I superfamily).* For this analysis, we first constructed a multiple sequence alignment of the aldolase I superfamily (CATH code: 3.20.20.70), using the program clustalw, and allowing no more than 10 percent gaps in the alignment. The resulting multiple sequence alignment is composed of 4,132 sequences of length 188 amino acids, and comprises 53 different enzyme functions at the finest-grained level of the EC hierarchy. For subsequent PCA (Jolliffe 2002), we encoded the sequences in the alignment as numeric strings (21 possible values per amino acid position, including gaps). The panels show the first two principal components (a) and the first and third components (b). The 53 different enzyme functions are color-coded according to the color bar to the right. Note the clear separation of some functions.

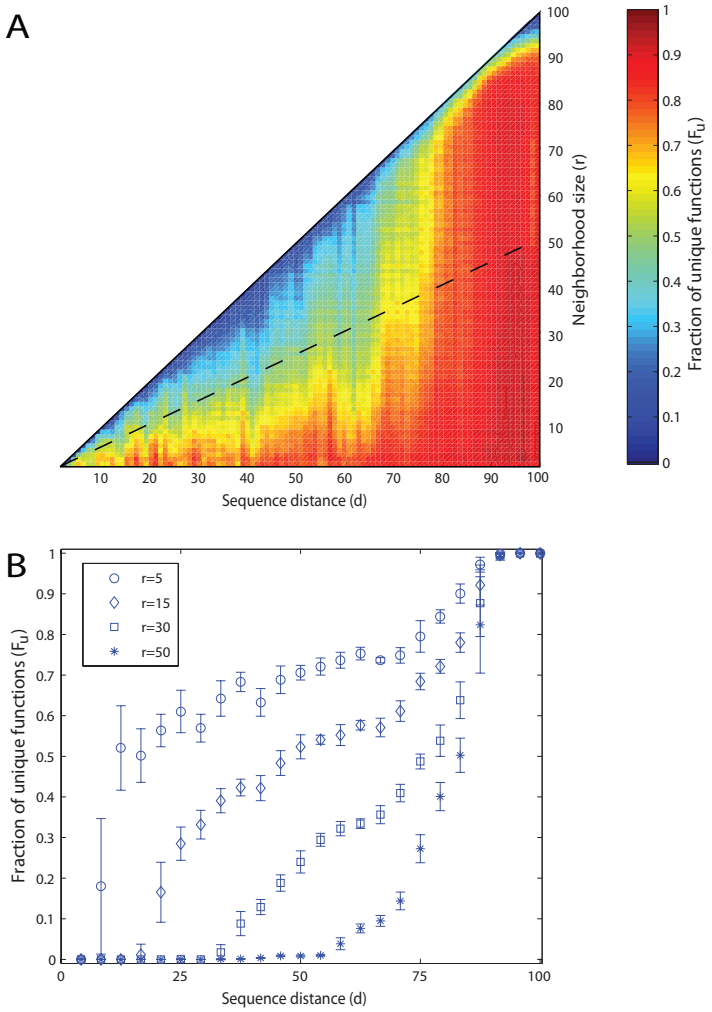


Figure S11. *Genotypic neighborhoods of proteins with a given structure.* The figure shows the dependency between the radius and distance of sequence neighborhoods, and the fraction F_u of functions unique to one neighborhood, for sequences folding into 36 different structures. The total set of multiple alignments we used in this analysis comprises a total of 18,117 sequences with lengths ranging from 100 to 400 amino acids, and spans 434 enzymatic functions covering all 6 EC classes. We analysed these sequences exhaustively. That is, for all possible pairwise sequence comparisons we computed their values of r, d and F_u. The heatmap shows F_u values at each combination of *d* and *r*, for the 26 structures (**a**) Heatmap of the fraction of unique functions (Fu) at different combinations of neighborhood radii (r) and sequences distances (d). (**b**) Fraction of unique functional F_u of unique functions versus sequence distance (expressed in percent) at a given neighborhood radius, as shown in the legend. Due to the sparsity of data, we grouped values into 20 different distance bins, each spanning d=5. Error bars represent standard errors calculated for these 20 bins. The CATH identifiers of the 36 superfamilies we used in this analysis

are listed here: 3.30.70.141; 3.30.420.10; 3.40.50.960; 2.70.40.10; 3.90.45.10; 3.40.50.2020; 3.20.19.10; 3.40.50.1470; 3.40.50.1360; 2.40.10.10; 3.90.1550.10; 3.90.226.10; 3.90.180.10; 3.40.50.880; 3.60.20.10; 3.40.50.620; 3.40.1210.10; 3.40.1160.10; 3.40.50.1240; 3.40.640.10; 3.60.15.10; 3.20.20.60; 3.20.20.70; 3.30.572.10; 3.90.550.10; 1.20.200.10; 3.40.1190.20; 3.30.930.10; 1.10.1040.10; 3.20.20.140; 3.40.50.1820; 3.20.20.210; 3.20.20.150; 3.40.718.10; 3.20.20.80; 1.10.630.10.

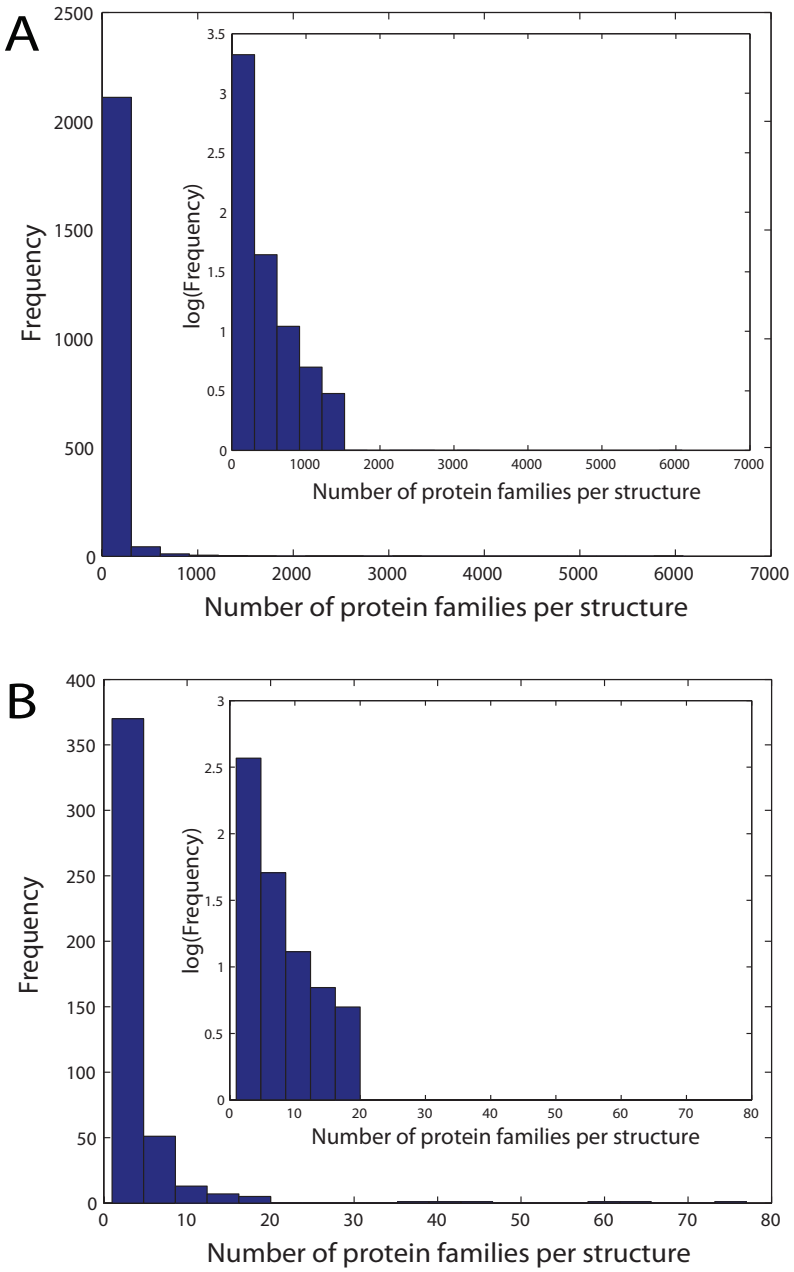


Figure S12. *Distribution of the number of protein families per structures.* (a) Distribution of the number of protein families per structure domain in the whole CATH database. This data is composed of 114,215 protein families grouped into 2,178 structures. (b) Distribution of the number of protein families per structure in our dataset composed of 39.529 sequences and 457 structures. More precisely, the notion of a protein family here corresponds to that of a CATH homologous superfamily (Greene et al, 2007). The insets show the same data, but with a $\log_{10}$-transformed vertical axis.

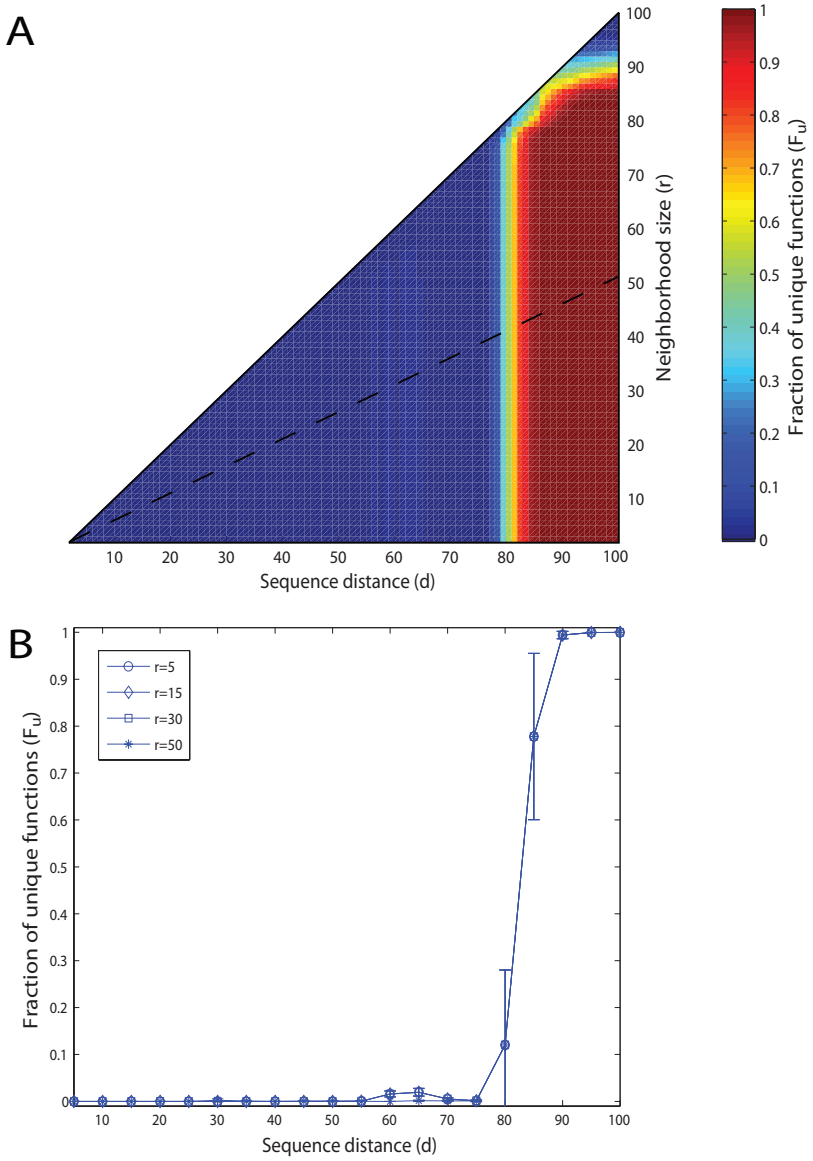


Figure S13. *Neighborhood diversity in functions depends on functionally versatile protein families and structures.* The figure shows the dependency between the radius and distance of two genotype neighborhoods, and the fraction F_u of functions unique to one neighborhood. (**a**) Heatmap of the fraction of unique functions (F_u) at different combinations of neighborhood radii (r) and sequences distances (d). The data is based on the major superfamily of the TIM barrel domain, aldolase I (CATH code: 3.20.20.70), which is composed of 4,132 sequences that carry out 53 different enzyme functions (see methods). These sequences can be grouped into 62 protein families. From this data set we selected the 30 protein families that carry out single enzyme functions. These families comprise 2,444 protein sequences and 27 enzyme functions. For all possible sequence pairs in this data set we computed values of d and F_u for different values of r. The heatmap shows F_u values over all distance-radius

combinations. (**b**) Fraction of unique functional variations versus sequence distance (expressed in percent) at constant neighborhood radii, as shown in the legend. Due to the sparsity of the data, we grouped values into 20 different distance bins, each spanning d=5. Error bars represent standard errors calculated for these 20 bins.

5. The organization of genotype space and the evolution of the protein genotype-phenotype map.

My goal in this chapter is threefold. First, to explore the role of dimensionality in the protein genotype space organization. Second, to show how this framework can be used to explain the distribution of neutral networks in genotype space, and additionally to study the evolution of the protein genotype-phenotype map. Finally I propose a simple model that aims to reconcile the extensive size variation observed in natural proteins with the genotype space concept.

5.1 The space of protein sequences

The number of known protein sequences may seem enormous but it only represents a tiny fraction of the theoretical number of possible sequences. If we consider that on average protein sequences possess 150 amino acids and are composed of an amino acid alphabet size of 20, then there may exist 20^{150} ($\sim 10^{196}$) possible combinations. Estimations of the total number of proteins per genome and the total number of genomes that remain unexplored suggest that we know much less than 1 percent of the total protein sequence space (Godzik 2011).

5.1.1 Subspaces, hyperspheres and hypersurfaces

In mathematical terms, any set of sequences in sequence space can be represented as a subgraph embedded in the *n*-cube, $Q^L_{|A|}$ (see Section 1.7.2). Formally, a *subgraph* or *subspace* embedded in $Q^L_{|A|}$ is a subset of vertices. The *size* of a subspace is the total number of its vertices. A subspace of at least two vertices is *connected* if, and only if, a path of edges exists between all pairs of vertices in it. As an example, the hypercube graph of Figure 1.9C is partitioned into three color-coded connected subspaces of unequal size (ie. 5, 5 and 6 vertices). The possible number of subgraphs embedded in the *n*-cube depends on *n*, and is generally very large (from Section 1.7.2 recall that $n=L(|\boldsymbol{A}|-1)$). More

precisely, there are $\binom{|\boldsymbol{A}|^L}{N}\sum_{i=0}^{E}\binom{\binom{N}{2}}{i}$ possible embeddings in $Q^L_{|A|}$ of a graph composed of N nodes and a maximum value of E edges.

Among all these possible subgraphs, and because of its mathematical simplicity, I focus on the *k*-ball. Although a *k*-ball may not be an accurate representation of a genotype network in sequence space, there are two reasons to use it as an approximation of a neutral network embedded in sequence space. First, empirical data on the protein universe and lattice models suggest that sets of related protein sequences cluster in isolated regions of sequence space. Second, the simplicity of this approximation allows us to study some properties of sequence space analytically.

As described above, the *k*-ball, $B^n_k(S_o)$, also called *k-neighborhood*, is a special case of a subspace, a discrete version of a *hypersphere* of radius *k* around a sequence S_o and embedded in a space of dimension $n=L(|\boldsymbol{A}|-1)$. The total number of sequences inside $B^n_k(x)$ is quantified by its volume $V_B(k)$ as:

$$V_B(k) = \sum_{i=0}^{i=k}\binom{L}{i}(|\boldsymbol{A}|-1)^i \tag{5.1}$$

Equation (5.1) simply condenses the binomial expansion:

$$1 + L(|\boldsymbol{A}|-1) + \frac{L(L-1)}{2}(|\boldsymbol{A}|-1)^2 + \ldots + \binom{L}{k}(|\boldsymbol{A}|-1)^k \tag{5.2}$$

where each term corresponds to the total sequences contained in the *i*-th shell of the *k-ball*. Therefore, the last term expresses the total area of the *k*-ball's surface or *k-surface*, $A_B(k)$:

$$A_B(k) = \binom{L}{k}(|\boldsymbol{A}|-1)^k \tag{5.3}$$

Figure 5.1 shows the fraction of the volume of an *n*-cube occupied by a *k*-ball embedded into it and with increasing radius *k*. Curves from left (blue) to right (red) represent different amino acid alphabets from 2 to 20. The length (*L*) is kept constant at *L*=100 (different lengths from *L*=25 on would yield qualitatively similar observations). For small *k,* the *k*-ball occupies a small fraction of the *n*-cube, until a sudden increase in the fractional volume occurs once *k* reaches the value of 1-κ (see Section 1.3). The explanation of this

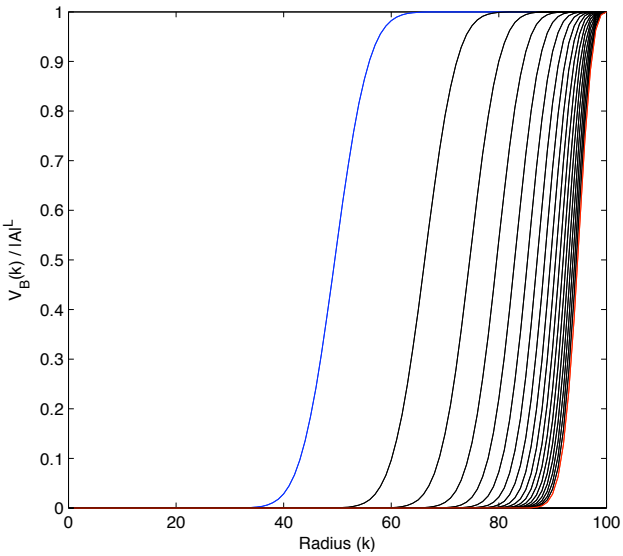


Figure 5.1. *Fraction of the n-cube ($Q^{100}_{|A|}$) occupied by the volume of a k-ball for L=100*. The alphabet size (|***A***|) ranges from 2 (blue) to 20 (red).

phenomenon is very simple and relies on the weighted binomial expansion displayed in Eq 5.2. *k*-balls of increasing size embedded into the *n*-cube accumulate sequences in different proportions.

In order to explore in more detail the structure of a *k*-ball, we calculate the fractional volume of a shell of the *k*-ball, $C^n_B(k)$:

$$C^n_B(k) = \frac{V_B(k) - V_B(k-1)}{V_B(L)} = \frac{A_B(k)}{|\boldsymbol{A}|^L} \tag{5.4}$$

Figure 5.2A shows $C^n_B(k)$ as a function of *k*. Most of the sequences tend to concentrate at $k \sim 1\text{-}\kappa$; which helps explain the sudden increase of the volume of the *k*-ball observed above (Figure 5.1).

Figure 5.2B shows the fraction of the *k*-ball volume occupied by the sequence of the outer shell (*k*-surface) of the *k*-ball at varying alphabet sizes and constant *L*=100. At small radii *k*, almost all of the sequences concentrate on the surface. This fraction falls rapidly as *k* increases, meaning that as shown before, *k*-balls do not concentrate sequences at their surface, but instead at distances that approximate the random threshold 1-κ. Taken together these observations show that the *k*-ball embedded in an *n*-cube concentrate sequences at distances that depend on the dimension *n* and particularly on the amino acid alphabet size.

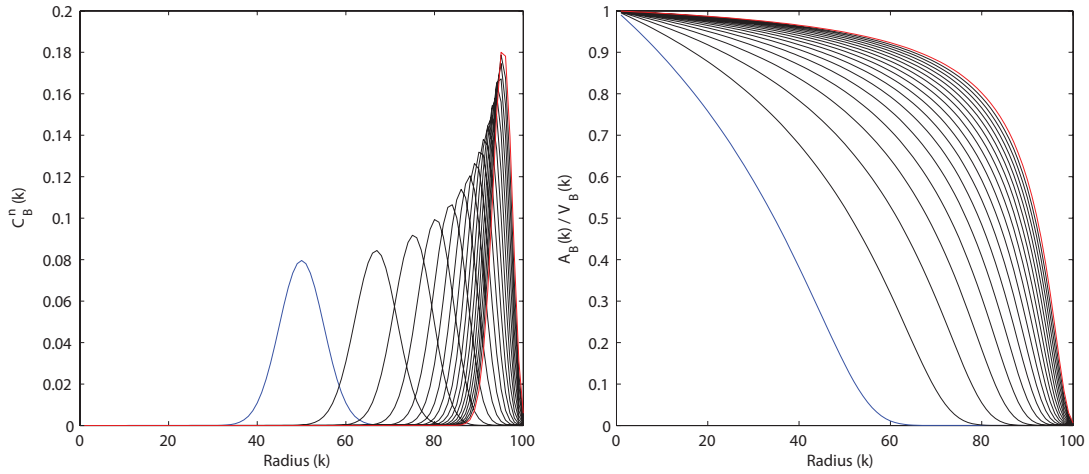


Figure 5.2. *Fractional volume of the k-ball*. Left. Fraction of the total *n*-cube volume occupied by the outer shell of the *k*-ball at increasing $|\boldsymbol{A}|$. Right. Fraction of the *k*-ball occupied by a *k*-surface. In both figures the alphabet size ($|\boldsymbol{A}|$) ranges from 2 (blue) to 20 (red) and sequence length is kept constant (L=100).

5.1.2 The distances to a k-surface.

To understand the organization of protein sequence space, the exploration of a single k-ball does not suffice. I next analyze the distance distribution between a particular sequence and sequences that compose the surface of a *k*-ball. This will provide information on the distance distribution between sequences that belong to different neigborhoods in sequence space.

Figure 5.3 illustrates this idea. Sequence S_1 (blue) is at a distance d from sequence S_o (red). We are interested in the distribution of distances between sequence S_1 and sequences contained in $A_B(k)$ (ie. the *k*-surface centered at sequence S_o). In order to explore this distribution, we note that Eq. 5.1 can be extended using the *Chu-Vandermonde identity* (Askey 1975):

$$\binom{m+n}{r} = \sum_{i}^{r} \binom{m}{i}\binom{n}{r-i} \tag{5.5}$$

where $m, n, r \in \mathrm{N_o}$. Using equations 5.3 and 5.5, we obtain:

$$A_B(k) = \sum_{j=0}^{k}\left[\binom{L-k}{j}(|\boldsymbol{A}|-1)^j\binom{k}{k-j}(|\boldsymbol{A}|-1)^{k-j}\right] \tag{5.6}$$

Eq. 5.6, like Eq. 5.3, enumerates all sequences at a distance k from a target sequence (S_o). However, Eq. 5.6 describes all possible ways in which k mutations may occur in S_o, and are present in $A_B(k)$. In this way, one can obtain the distribution $\boldsymbol{D}$ of distances from S_1 to all sequences in $A_B(k)$ (see Figure 5.3).

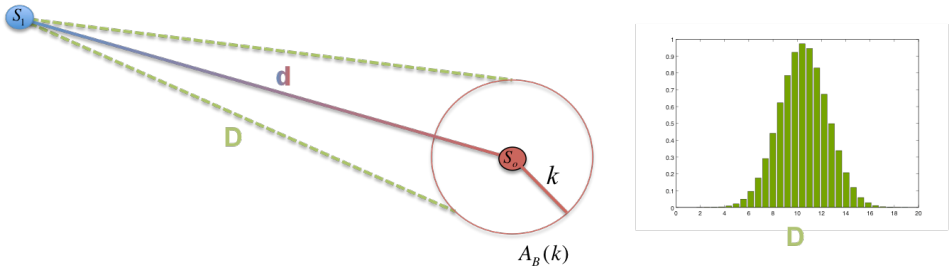


Figure 5.3. *Distance distributions in sequence space.* Hypothetical genotypes S_1 (blue) and S_o (red) are at a distance d in genotype space. The k-surface of genotype S_o, $A_B(k)$, corresponds to all genotypes at distance k from S_o, and is represented as a red circle. The distribution of distances from S_1 to any sequence in $A_B(k)$ is shown as the hypothetical distance distribution $\boldsymbol{D}$ (green) on the right.

Eq. 5.6 can also be applied to several other problems related to the distribution and organization of protein genotype networks in sequence space.

I used Eq. 5.6 to compute the distance distribution $\boldsymbol{D}$ for different radii k and for all possible distances $d(S_1,S_o)$, as shown in Figure 5.4 for four values of k and a sequence composed of a binary alphabet ($|\boldsymbol{A}|$=2) and L=25. Each distance distribution $\boldsymbol{D}$ for a given $d(S_1,S_o)$, is represented using a single color and was normalized to a total area of 1.0. Small radii (k=2 and k=5, 8 and 20 percent of L, respectively) produce sharply peaked distributions that separate in sequence space proportionally to $d(S_1,S_o)$ (Figure 5.4, upper panels, k=2 and k=5). Similarly, at large radii (k=20, 80 percent of L), distances decrease as $d(S_1,S_o)$ increases (Figure 5.4, panel k=20). This is because at larger $d(S_1,S_o)$, S_1 approaches the set $A_B(k)$ centered around S_o. For k=d, $\boldsymbol{D}$ becomes the distance distribution between sequences contained in the k-surface, $A_B(k)$. At values of k close to $L/2$ (k=10, 40 percent of L), and an alphabet size of $|\boldsymbol{A}|$=2, $\boldsymbol{D}$ becomes independent of the distance between the targeted genotypes, $d(S_1,S_o)$. These observations provide a geometric interpretation of the random threshold of sequence divergence (1-κ).

5.1.3 The mean sequence divergence of a k-neighborhood.

An important biological property of a protein family in particular, and of sequence space in general, is the mean sequence divergence of a related set of sequences. This value corresponds to the mean sequence variation between sequences inside a k-ball, $B_k^n(S_o)$.

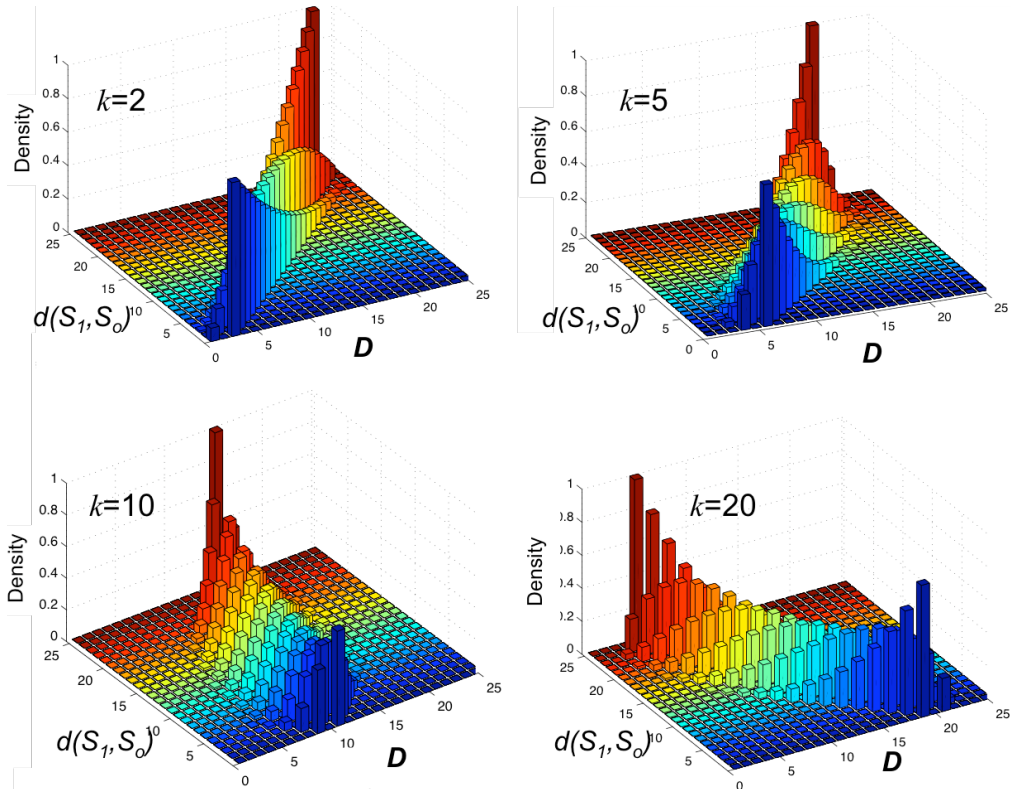


Figure 5.4. *Distribution of sequence distances for different k-surfaces*. Distance distributions (***D***) are computed for different k and $d(S_1,S_o)$ using equation 5.6. Values of the distributions are normalized to 1.0.

Eq. 5.6 can be used to explore sequence variation inside a k-ball by exploring sequence distributions at $d(S_1,S_o) \leq k$ (Figure 5.3). Figure 5.5 shows the percentage of mean pairwise sequence divergence calculated over all sequences that belong to a k-ball at different alphabet sizes that range between 2 to 20, for L=100. For small values of k, mean pairwise sequence divergence increases almost linearly with a slope of $\sim 2k$. The maximum value of mean pairwise sequence divergence equals approximately 50 percent, and does not vary strongly with alphabet size. For $|\boldsymbol{A}| = 2$, divergence reaches a maximum of 50 percent at 1-κ and remains constant at larger values of k. For larger alphabet sizes, however, and for $k >$ 50 percent, mean pairwise sequence divergence decreases almost lineraly at $k \sim$ 1-κ, and then remains constant at higher values of k. Although in general we are interested in the mean sequence divergence of small k-balls ($k <$ 50 percent), these observations show another property of genotype space that results from its multiple dimensions.

The data in figure 5.5 indicate that due to the intrinsic geometry of the n-cube mean pairwise sequence divergence increases only up to k-ball radii of 50

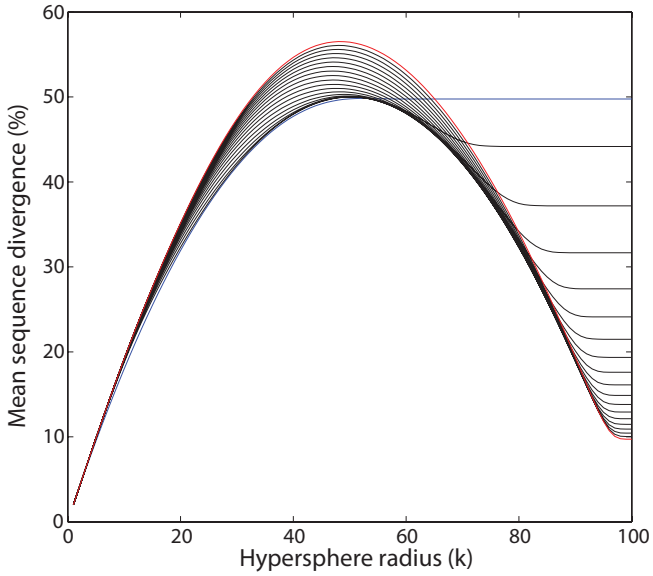


Figure 5.5. *Mean sequence divergence of a k-neighborhood*. Total mean sequence divergence is calculated for a k-ball of alphabet sizes ranging from 2 (blue) to 20 (red). Length is kept constant at L=100.

percent of the space size have been reached. This phenomenon is independent of sequence space dimension. At $k > L/2$, the k-ball 'folds onto itself', meaning that sequences contribute negatively to mean pairwise sequence divergence. This phenomenon continues until $k > 1-\kappa$, then mean pairwise sequence divergence remains constant.

5.2 Prototype sequences distribute randomly in sequence space.

Now that I have discussed some features of the organization of sequence space, I will use data from a 2D HP lattice model of length 25 to study the distribution of neutral networks in sequence space. Table 5.1 and Figure 5.6A shows some of the statistics of this model. The number of sequences per conformation shows a non-uniform distribution. Out of the 2^{25} possible sequences only 2 percent fold into a single conformation with a unique minimum energy. A fraction of 6 percent of the total foldable sequences is completely isolated in space, that is, they have neutral networks with size 1. The remaining 94 percent of foldable sequences fall into neutral networks of variable size. The largest network has 326 sequences and on average, neutral networks of size greater than one comprise 5 sequences. All genotypes with the same fold form a genotype set. 71 percent of the genotype sets consist of a single neutral network

Table 5.1. *Statistics of the 2D HP lattice model (L=25).*

Object	Total
Total sequences	33,554,432
Foldable sequences	765,147
Neutral networks (size=1)	50,407
Neutral networks (size>1)	97,847
Total neutral networks	148,254
Genotype sets (size=1)	76,800
Genotype sets (size>1)	30,536
Total conformations (genotype sets)	107,336

The largest genotype set is composed of nine neutral networks. Figure 5.6 shows the mean and maximum distances of sequences in a neutral network. The average and maximum mean sequence distance are 1.6 and 5.7; respectively. The maximum sequence distance distribution shows average and maximal values of 1.7 and 15, respectively.

In order to explore the distribution of neutral networks in sequence space, I identified the prototype sequence of each neutral network (network size > 1; 97,847 networks, see Table 5.1) in the data by using *single-linkage clustering* (Everitt 2011) based on all pairwise distances of sequences in a network. In the case of neutral networks of size one, I considered the single sequence as the prototype. To explore the distribution of prototype sequences in genotype space, I calculated the distance distribution from a prototype sequence to all other prototype sequences in the data set. In order to explore whether there is any

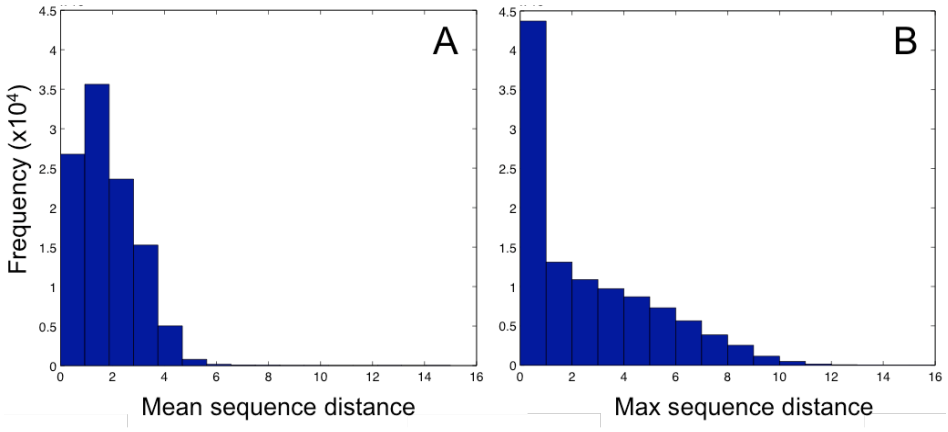


Figure 5.6. *Mean and maximum sequence distances in the 2D HP model L=25.* A. Distribution of the mean sequence distance between sequences that belong to the same neutral network. B. Maximum sequence distance observed per neutral network.

influence of network size on the distribution of distance, I selected three groups of sequences, composed of 10^3 prototype sequences each. The first comprised, a random sample of prototype sequences of neutral networks of size one. The second comprised 10^3 prototype sequences of the largest neutral networks in the data set. A third group comprised 10^3 random sequences that may or may not be uniquely foldable. For each sequence in each of these sets I calculated a distance distribution to all other prototype sequences in the complete data set, that is, regardless the size of the networks they belong to. These distributions are shown in Figure 5.7A. A chi-square goodness-of-fit test shows that there are no

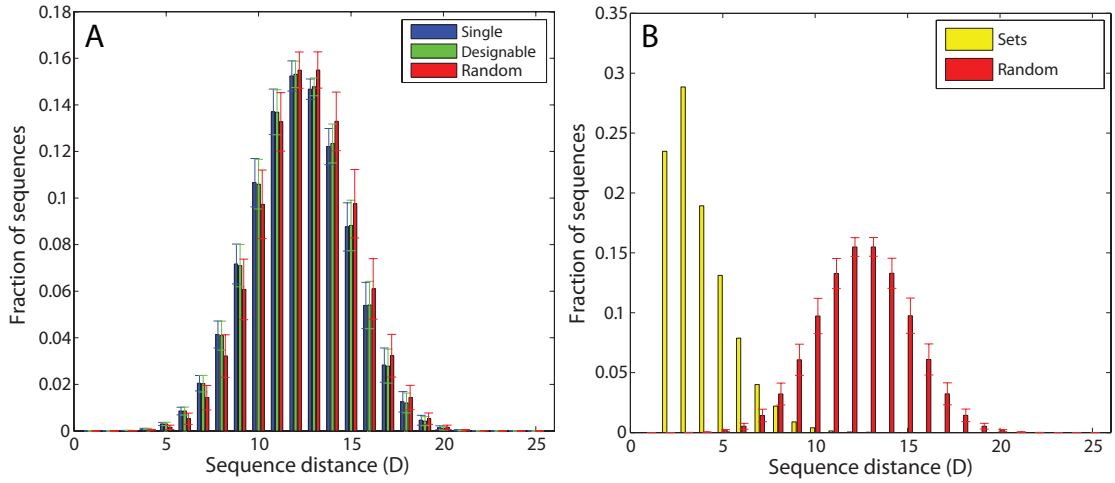


Figure 5.7. *Distances between prototype sequences distribute randomly in genotype space, whereas prototype sequences of the same genotype set do not.* **A.** Distance distributions between prototype and random sequences in the HP lattice model L=25. Each distribution was obtained by comparing the 10^3 sequences in each data set to all prototype sequences in the complete data set. The 'single' set, composed of prototype sequences of 10^3 randomly chosen neutral networks of size 1, is shown in blue. The 'designable' set, composed of prototype sequences of the 10^3 most designable neutral networks in the data set is shown in green. The 'random' set, composed of 10^3 randomly chosen sequences among the 2^{25} possible HP sequences of size L=25 is shown in red. A pearson χ^2 goodness-of-fit test shows no significant statistical difference neither between pairs of these distributions nor between any of the distributions and a normal distribution (χ^2, p-value $<10^{-16}$). Error bars show one standard deviation from the mean. B. Distance distributions between prototype sequences that belong to the same genotype set. In yellow, distance distributions between prototype sequences of the same genotype set. In red, distances between randomly chosen sequences. The number of pairwise distance comparisons is the same as for the distribution in yellow. A Pearson χ^2 goodness-of-fit test shows no significant statistical difference between these two distributions. Error bars show one standard deviation from the mean.

significant differences neither between the distance distribution of the three sets nor between any of them and a normal distribution (χ^2, p-value < 10^{-16}). In other words, pairwise distances between prototype sequences distribute as expected by chance. Furthermore, this result shows that these distributions are independent of network size.

Similarly, we may ask for the distribution of pairwise distances between prototype sequences of neutral networks that belong to the same genotype set. Figure 5.7B shows that prototype sequences that fold into the same conformation are closer in genotype space than expected by chance.

Taken together these results suggest that the distribution of prototype sequences in genotype space is determined by the intrinsic organization of genotype space itself. Specifically, sequences in a genotype space with $|\boldsymbol{A}|$=2 and L=25 concentrate predominantly at distances of ~12-point mutations (Figure 5.2A). Similarly, in spaces of larger alphabet size, it is expected that sequences concentrate at distances values of 1-κ. Since the majority of neutral networks are very small, we conclude that neutral networks distribute randomly in genotype space at distances proportional to 1-κ.

5.3 The evolution of the protein genotype-phenotype map.

Empirical evidence exists that supports the idea of the evolution of the genetic code by the addition of new amino acids (Hatfield and Gladyshev 2002; Hendrickson et al 2004; Lu and Freeland 2006). The first direct consequence of such progressive addition of new amino acids was the increase in the dimensionality of genotype space.

Using protein simple exact models the effect of dimensionality on the genotype-phenotype map can be readily tested. A previous study explored the impact of alphabet sizes 2, 4 and 20 amino acids on the designability of lattice proteins (Buchler and Goldstein 1999a; 1999b; 2000). This study showed that the distribution of neutral network size is affected by amino acid alphabet size. An exponential distribution for small alphabets ($|\boldsymbol{A}|$=2, $|\boldsymbol{A}|$=4), turns into a Gaussian distribution at larger alphabets ($|\boldsymbol{A}|$=20). Strikingly, designable conformations at lower alphabet sizes were not necessarily designable at higher alphabet sizes (Buchler and Goldstein 1999). In addition to influence protein

designability, theory suggests that two properties of the amino acid alphabet, namely size and chemical diversity, may have allowed the appearance of energy minima that expanded the available conformational space of proteins (Wolynes 1997).

As shown in section 5.1.1, genotype spaces of higher dimensions concentrate sequences at larger distances and therefore may influence the sequence-structure map of proteins. Since pairwise distances between prototype sequences of neutral networks distribute randomly in genotype space, we speculate that higher alphabet sizes may have allowed the appearance of new and larger neutral networks.

A thorough exploration of the effect of space dimensionality on the protein sequence-structure map will shed light on different aspects of protein evolution. First, it may advance our current view on the evolution of the genetic code and the space of protein structures. Second, in protein engineering, it may help design new conformations by empirical exploration of favorable sequence space dimensions. Third and more importantly, the analysis of the effects of genotype space dimension on the sequence–structure map may provide insights into the evolution of evolvability.

5.4 Genotype space and protein size: a simple model.

Natural proteins show extensive variation in their size, so a definition of genotype space purely based on a single dimension, may not capture the real features of the protein universe, nor help explain its evolution. Additionally, an evolutionary model based purely on structural space loses the advantages of describing genetic population parameters such as evolutionary rate (eg. the rate of mutation or indels). This information might be an important part of a unified view of protein evolution.

Here I propose a simple model to alleviate the problem of the different dimensions observed in the protein universe by unifying both, the sequence and the structural frameworks. In this augmented model of protein sequence space, the universe of proteins is embedded in different spaces of increasing dimensions (Figure 5.8). Each genotype space represents a range of sequence length whose dimensions allow them to be compared. Folding studies suggest

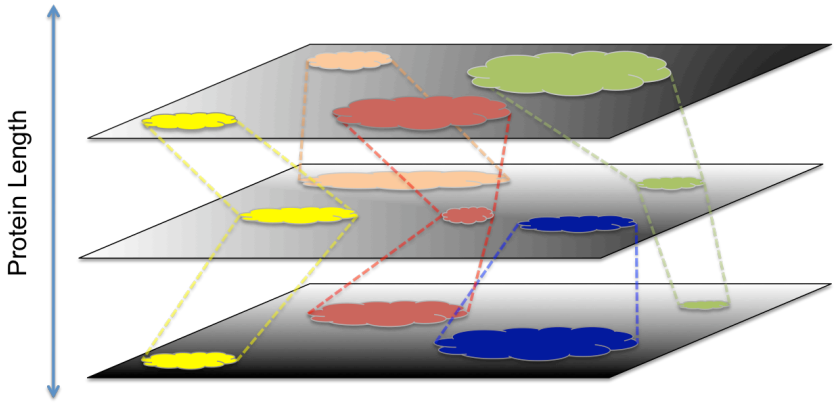


Figure 5.8. *An augmented model of the protein sequence space*. Each flat surface represent a sequence space of a given dimension. Contiguous spaces have similar dimensions. Colored clouds represent neutral networks. Dashed lines show the possible continuity between folds' neutral networks through different spaces.

that the same protein fold may be found in sequences that can vary in length up to 50 amino acids (Craik et al 1983; Bashford et al 1987). In the model, protein sequences that evolve by point mutations keep their size constant and therefore evolve into the same genotype space. In contrast, mechanisms such as insertions, deletions or heterologous recombination, can cause a considerable change in protein size and promote the transition from one space to another.

Once a sequence fragment has been inserted into a fold, there are three possible outcomes. First, the fold can be disrupted in such a way that the resulting protein does not fold. Second, the insertion can cause no effect on the original fold and therefore the neutral network expands into the adjacent genotype space of higher dimension. Third, the insertion may give rise to a new fold and therefore a new neutral network in the adjacent genotype space of higher dimension, arises. As a consequence, the neutral network of a given fold may inhabit only some spaces and not others (blue networks in Figure 5.8). Other networks may preferably inhabit a particular space and show a reduced size in adjacent spaces (orange network in Figure 5.8). In contrast, certain folds may be very resistant to changes in their size and therefore be equally present across different spaces (yellow network in Figure 5.8).

6. Conclusions

The sequence – structure map of RNA and proteins are ideal systems to explore questions related to evolutionary biology, for two main reasons. First, macromolecules are the simple among the complex. They constitute both genotype and phenotype in a single entity and the rules that govern their phenotypes can be derived from physicochemical principles. Second, we possess both simple polymer models and sufficient empirical data from sequence and structures that allow the exploration of their properties systematically. In this dissertation I have explored the organization of RNA and protein genotype – phenotype maps with emphasis on proteins.

The study based on simple biopolymer models (Chaper 2) showed two well-known commonalities of the relationship between genotype (sequence) and phenotype (structure) of RNA and proteins. First, many phenotypes are formed by more than one genotype. The genotypes adopting any one phenotype usually form connected networks of genotypes. Second, some phenotypes are adopted by many more phenotypes than others. The RNA and protein genotype-phenotype relationships also show major differences. The first of them is that only a small fraction of protein genotypes adopts a unique fold. This is not the case for RNA, where most genotypes adopt a unique fold.

A second major difference is that model proteins form many more structures – even though fewer of their sequences fold – than RNA molecules. This property probably arises from the larger number of possible contacts that each monomer can have in a protein. Three more differences between RNA and protein follow from the first two differences: The number of genotypes that form a specific phenotype is smaller for proteins, the number of genotypes in any one genotype network is also smaller for proteins; and the average and maximum distances of genotypes with the same phenotype are smaller for proteins.

A last and important final difference regards shape space covering (Gruener et al 1996b). A ball of a given radius around an RNA molecule in sequence space contains a larger percentage of phenotypes than a ball of the same radius around a protein molecule. Shape space covering indicates that genotype networks are highly interwoven in the case of RNA (Schuster 1994),

and less so in the case of proteins (Bornberg-Bauer and Chan 1999). This last difference has tentative support from comparative analyses of natural RNA and protein molecules. In general, data presented in this dissertation show that RNA genotype space is more conducive to evolutionary searches for novel structures than proteins.

In this dissertation I have also explored the relation between structure and function (Chapter 3). I have shown that highly designable proteins evolve more functional innovations on large evolutionary time-scales. Two measures of designability estimate a given domain's ability to explore sequence space and access a diverse spectrum of functions. Because functional diversity is a record of past evolutionary innovations, this means that more designable proteins may have a greater facility to evolve new functions. In addition, because proteins of similar structure are connected in genotype space (Babajide et al. 1997, 2001; Bornberg-Bauer 1997; Bastolla et al. 1999; Wroe et al. 2007), more robust proteins may show greater propensity to evolve functional innovations.

The association between protein robustness and innovation holds for two complementary measures of functional diversity: diversity of enzymatic functions and gene ontology-based diversity of molecular functions. It also holds for two different measures of designability: one based purely on structural information, and the other based on the number of sequences associated with each protein fold.

Complex relationships with other variables notwithstanding, it is clear that designable and robust proteins have evolved many novel functions. This shows that a pattern derived from recent experimental findings, and applicable only to laboratory time-scales, also holds on vastly greater geological time-scales (Aharoni et al. 2005; Bloom et al. 2006). The possible explanation has its root in how populations explore vast sequence spaces: populations of highly robust folds can explore sequence space rapidly, and thus access large amounts of structural diversity in their neighborhood (Sumedha et al 2007). A small fraction of this diversity can subsequently give rise to proteins with new functions.

In a third study (Chapter 4), I explored the organization of functions in genotype space. I observed that new functions are encountered at varying sequence distances as proteins diverge in sequence space, and that this property

can be attributed to the fact that some protein families perform multiple functions. While for short distances in genotype space this diversity is moderate, it increases at larger distances and once the structure conservation threshold is crossed (Rost 1999), we observed an explosion in the accessibility of new structures, and consequently an enormous increase in functional diversity.

I pointed out three important observations in this study. First, different functions are carried out by different numbers of sequences and structures. Second, most functions are restricted to single structures, but some can be carried out by many structures. Relatedly, most protein families are associated with only one function, as was also shown previously based on fewer data (Todd et al 2001). Third, and most important, different genotype neighborhoods tend to contain a different spectrum of functions, whose diversity increases with increasing distance of these neighborhoods in sequence space. This observation has obvious implications for the evolution of novel protein functions. That is, by exploring a genotype network, proteins can explore ever-changing sequence neighborhoods, and an ever-changing spectrum of novel enzymatic functions. Fourth, the number of structures per function has a nonuniform distribution, even after controlling for the number of known sequences for each structure. This observation hints that some functions may indeed be more abundant in sequence space than others.

The phenotypic diversity of different neighborhoods in sequence space (Chapter 4) also has a flip side: It means that not all protein functions occur in every neighborhood of sequence space. In other words, the evolution of novel protein functions is constrained by an individual or a population's location in sequence space. A consequence of such constraints is evolutionary stasis, where genotypes but not phenotypes in a population change while the population explores a genotype network. Such stasis is interrupted by the discovery of novel phenotypes when a population arrives at a neighborhood where such novel phenotypes are found. The very feature that both facilitates evolutionary exploration of novel functions and causes their constrained evolution is probably a generic property of protein sequence space.

In the last section of this dissertation I have shown that the organization of genotype space is determined by its dimension, which is defined by two properties of sequences, length and monomer alphabet size. The intrinsic organization of protein genotype space and properties of protein neutral networks revealed by simple models, support the sequence–structure–function relationship observed in empirical data.

Analyses of the organization of genotype space in terms of the pairwise distances between sequences support both the structural diversity observed in the Chothia-Lesk plot (see Figure 1.8) and the consequent rise of functional diversity (see Figures 3 and 4 in Chapter 4) at distances that approximate the random threshold (1-κ).

Previous studies have suggested that neutral networks concentrate in certain regions of genotype space and therefore do not distribute randomly in sequence space (Mann 2011). However, in this dissertation I have shown that when the underlying organization of genotype space is taken into account, pairwise distances between prototype sequences resemble the pairwise distance distribution of random sequences in genotype space.

The study of the organization of genotype space also sheds light on the evolution of the protein genotype-phenotype map and supports previous results from simulation studies. As new amino acids were incorporated into the genetic code, the dimension of genotype space increased accordingly, which had three main consequences. First, the number of phenotypes increased. A larger number of amino acids allowed the stabilization of structures with higher energy minima and as a result, new structures arose (Wolynes 1997). Second, in genotype spaces of increasing dimensions neutral networks can grow larger. Simulation studies using protein lattices with different alphabet size show that this is the case (Buchler and Goldstein 1999). Third, as predicted by the organization of sequences in genotype space, at larger alphabet sizes structural and functional diversity accumulates further away from each sequence. Therefore genotype spaces composed of smaller alphabet size may contain more evolvable proteins. This may be one of the reasons for the success of reduced amino acid alphabets in protein design experiments (Riddle et al 1997; Dokholyan 2004). Taken together, these observations suggest that structural and functional diversity are

expected to be rarer in genotype spaces of larger dimensions. However, this phenomenon may be counteracted by the enlargement of neutral networks in such spaces. Simulation studies using protein simple exact models of different dimensions, as well as the analysis of protein folds, may shed further light on these conjectures.

Bibliography

1. Abkevich VI, Gutin AM and Shakhnovich EI. Specific Nucleus as the Transition State for Protein Folding: Evidence from the Lattice Model. *Biochemistry* 33:10026–10036 (1994).
2. Aharoni A, Gaidukov L, Khersonsky O, Gould SMcQ, Roodveldt C and Tawfik DS. The evolvability of promiscuous protein functions. *Nature Genetics* 37:73-76 (2005).
3. Aiello B and Leighton T. Coding theory, hypercube embeddings, and fault tolerance. *Proceeding SPAA '91 Proceedings of the third annual ACM symposium on Parallel algorithms and architectures*. Pag 125-136. New York, NY, USA (1991).
4. Alberch P. From genes to phenotype: dynamical systems and evolvability. *Genetica* 84:5-11 (1991).
5. Altenberg L. The evolution of evolvability in genetic programming, In: K. Kinnear, Editor, *Advances in Genetic Programming,* 47-74 MIT Press.
6. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol* 215:403–410 (1990).
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402 (1997).
8. Anfinsen CB. Principles that govern the folding of protein chains. *Science* 81:223-230 (1973).
9. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM and InterPro Consortium. InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16:1145-1150 (2000).
10. Askey R. *Orthogonal polynomials and special functions*. Philadelphia. Society for industrial and applied mathematics (1975).
11. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS and Stadler PF. Exploring protein sequence space using knowledge-based potentials. *J Theor Biol* 212:35-46 (2001).
12. Babajide A, Hofacker IL, Sippl MJ and Stadler PF. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des* 2:261-269 (1997).
13. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 28: 304–305 (2000).
14. Baker D. A surprising simplicity to protein folding. *Nature* 405:39-42 (2000).
15. Bashford D, Chothia C and Lesk AM. Determinants of a protein fold: unique features of the globin amino acid sequences. *J Mol Biol* 196:199-216 (1987).
16. Bashton M and Chothia C. The generation of new protein functions by the combination of domains. *Structure* 15:85-99 (2007).
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. *Nucleic Acids Res* 28: 235–242 (2000).

18. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L, Kopp J, Podvinec M, Adams PD, Carter LG, Minor W, Nair R and La Baer J. The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res* 37:D365-D368 (2009).
19. Bloom JD and Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci USA* 106:9995-10000 (2009).
20. Bloom JD, Raval A and Wilke CO. Thermodynamics of neutral protein evolution. *Genetics* 175:255-266 (2007).
21. Bloom JD, Drummond DA, Arnold FH and Wilke CO. Structural Determinants of the Rate of Protein Evolution in Yeast. *Mol Biol Evol* 23:1751-1761 (2006).
22. Bloom JD, Romero PA, Lu Z and Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol Direct* 2:17 (2007).
23. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA* 102:606-611 (2005).
24. Bloom JD, Labthavikul ST, Otey CR and Arnold FH. Protein stability promotes evolvability. *Proc Natl Acad Sci* 103:5869-5874 (2006).
25. Bloom JD, Wilke CO, Arnold FH, Adami C. Stability and the evolvability of function in a model protein. *Biophys J* 86:2758-2764 (2004).
26. Bornberg-Bauer E. How are model protein structures distributed in sequence space? *Biophys J* 73:2393-2403 (1997).
27. Bornberg-Bauer E and Chan HS. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA* 96:10689-10694 (1999).
28. Bornberg-Bauer E, Huylmans AK and Sikosek T. How do new proteins arise? *Curr Opin Struct Biol* 20:390-396 (2010).
29. Branden C and Tooze J. *Introduction to protein structure*. New York, Garland (1999).
30. Bryngelson JD, Onuchic JN, Socci ND and Wolynes PG. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins* 21:167-195 (1995).
31. Buchler NEG and Goldstein RA. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins* 34:113-124 (1999a).
32. Buchler NEG and Goldstein RA. Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *Journal of Chemical Physics* 111:6599-6609 (1999b).
33. Buchler NEG and Goldstein RA. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: A consensus. *Journal of Chemical Physics* 112:2533-2547 (2000).
34. Chan HS and Bornberg-Bauer E. Perspective on protein evolution from simple exact models. *Appl Bioinformatics* 1:121-44 (2002).

35. Chen T, Vernazobres D, Yomo T, Bornberg-Bauer E and Chan HS. Evolvability and single-genotype fluctuation in phenotypic properties: a simple heteropolymer model. *Biophys J* 98:2487-2496 (2010).
36. Chen Y, Shoichert B and Bonnet R. Structure, function, and inhibition along the reaction coordinate of CTX-M beta-lactamases. *J Am Chem Soc* 127:5423-5434 (2005).
37. Choi IG and Kim SH. Evolution of protein structural classes and protein sequence–families. *Proc Natl Acad Sci USA* 103:14056-14061 (2006).
38. Chothia C. Proteins. One thousand families for the molecular biologist. *Nature* 357:543-544 (1992).
39. Chothia C and Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823-826 (1986).
40. Chothia C, Gough J, Vogel C and Teichmann SA. Evolution of the protein repertoire. *Science* 300:1701-1703 (2003).
41. Chubb D, Jefferys BR, Sternberg MJE and Kelley LA. Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics* 26:2664-2671 (2010).
42. Chung SY and Subbiah S. A structural explanation for the twilight zone of protein sequence homology. *Structure* 4:1123-1127 (1996).
43. Conant GC and Wolfe KH. Turning a hobby into a job: how duplicate genes find new functions. *Nat Rev Genet* 9:938-950 (2008).
44. Coulson AF and Moult JA. Unifold, mesofold, and superfold model of protein fold use. *Proteins* 46:61-71 (2002).
45. Craik CS, Rutter Wj and Fletterick R. Splice junctions: association with variation in protein structure. *Science* 220:1125-1129 (1983).
46. Crick F. Central dogma of molecular biology. *Nature* 227:561-563 (1970).
47. Crippen GM and Chhajer M. Lattice models of protein folding permitting disordered native states. *Journal of Chemical Physics* 116:2261-2268 (2002).
48. Cui Y, Wong WH, Bornberg-Bauer E and Chan HS. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA* 99:809-814 (2002).
49. Dawkins R. *The evolution of evolvability*. In: Artificial life, C. Langton (ed.), Addison Wesley (1989).
50. Dayhoff MO. The origin and evolution of protein superfamilies. *Fed Proc* 35:2132–2138 (1976).
51. Dayhoff MO and Schwartz RM. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (chapter 22) (1978).
52. de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, Gibson G, Hansen TF, Krakauer D, Lewontin RC, Ofria C, Rice SH, von Dassow G, Wagner A and Whitlock MC. Perspective: Evolution and detection of genetic robustness. *Evolution* 57:1959-1972 (2003).
53. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B and Orengo C. PSI-2: structural genomics to cover protein domain family space. *Structure* 17:869-881 (2009).
54. Devos D and Valencia A. Practical limits of function prediction. *Proteins* 41:98-107 (2000).

55. Dias CL and Grant M. Designable proteins are easy to unfold. *Phys Rev E* 74:042902 (2006).
56. Dill KA and Chan HS. From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10-19 (1997).
57. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD and Chan HS. Principles of protein folding – a perspective from simple exact models. *Protein science* 4:561-602 (1995).
58. Dokholyan NV. What is the protein design alphabet? *Proteins* 54:622-628 (2004).
59. Doolittle WF. Lateral genomics. *Trends Cell Biol* 9:M5–M8 (1999).
60. Drummond DA and Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341-352 (2008).
61. Drummond DA and Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10:715-724 (2009).
62. Drummond DA, Silberg JJ, Meyer MM, Wilke CO and Arnold FH. On the conservative nature of intragenic recombination. *Proc Natl Acad Sci USA* 102:5380–5385 (2005).
63. England JL and Shakhnovich EI. Structural Determinant of Protein Designability. *Phys Rev Lett* 90:218101 (2003).
64. Erdin S, Lisewski AM and Lichtarge O. Protein function prediction: towards integration of similarity metrics. *Current Opinion in Structural Biology* 21:180-188 (2011).
65. Everitt BS. *Cluster analysis*. 5th Ed. Chichester, Wiley (2011).
66. Ferrada E and Wagner A. Protein robustness promote evolutionary innovation on large evolutionary time-scales. *Proc Biol Sci Lond B* 275:1595-1602 (2008).
67. Ferrada E and Wagner A. Evolutionary Innovations and the Organization of Protein Functions in Genotype Space. *PLoS ONE* 5: e14172 (2010).
68. Ferrada E and Wagner A. A comparison of genotype-phenotype maps for RNA and proteins. (*Submitted*).
69. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR and Bateman A. The Pfam protein families database. *Nucleic Acids Res* 38:D211-222 (2010).
70. Frauenfelder H, Sligar SG and Wolynes PG. The energy landscapes and motions of proteins. *Science* 254:1598-1603 (1991).
71. Gavrilets S. Fitness Landscapes and the Origin of Species, *Monographs in Population Biology*; v. 41. Princeton University Press, Princeton, N.J., Oxford, England (2004).
72. Geer Y, Domrachev M, Lipman DL, Bryant SH. CDART: protein homology by domain architecture. *Genome Research* 12:1619-1623 (2002).
73. George A and Wilson WW. Predicting protein crystallization from a dilute solution property. *Acta Cryst* D50:361-365 (1994).
74. Gilbert W. Why genes in pieces? *Nature* 271:501 (1978).
75. Giugliarellia G, Micheletti C, Banavar JR and Maritan A. Compactness, aggregation, and prionlike behavior of protein: A lattice model study. *Journal of Chemical Physics* 113:5072-5077 (2000).

76. Godzik A. Metagenomics and the protein universe. *Current Opinion in Structural Biology* 21:398-403 (2011).
77. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396-1407 (2010).
78. Govindarajan S and Goldstein RA. The foldability landscape of model proteins. *Biopolymers* 42:427-438 (1997).
79. Govindarajan S, Recabarren R and Goldstein RA. Estimating the total number of protein folds. *Proteins* 35:408–414 (1999).
80. Grishin NV. Fold change in evolution of protein structures. *Journal of Structural Biology* 134:167-185 (2001).
81. Hamming RW. *Coding and Information Theory*. Prentice Hall, Englewook Cliffs, N.J (1980).
82. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68:669-685 (2004).
83. Hartling J and Kim J. Mutational robustness and geometrical form in protein structures. *J Exp Zool (Mol Dev Evol)* 310B:216–226 (2008).
84. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the genetic code. *Mol Cell Biol* 22:3565-3576 (2002).
85. Hegyi H and Gerstein M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* 11:1632-1640 (2001).
86. Helling R, Li H, Mélin R, Miller J, Wingreen N, Zeng C and Tang C. The designability of protein structures. *J Mol Graph Model* 19:157-167 (2001).
87. Hendrickson TL, de Crecy-Lagard V, Schimmel P. Incorporation of nonnatural amino acids into proteins. *Annu Rev Biochem* 73:147-176 (2004).
88. Hirst JD. The evolutionary landscape of functional model proteins. *Protein Eng* 12: 721-726 (1999).
89. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M and Schuster P. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh Chem* 125: 167-188 (1994).
90. Holm L and Sander C. Mapping the protein universe. *Science* 273:595-603 (1996).
91. Hubbard SJ, Eisenmenger F and Thornton JM. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci* 3:757-768 (1994).
92. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH and Yeats C. InterPro: the integrative protein signature database. *Nucleic Acids Res* 37:D224-228 (2009).
93. Irbäck A and Sandelin E. On Hydrophobicity correlations in protein chains. *Biophys J* 79: 2252-2258 (2002).
94. Irbäck A and Troein C. Enumerating Designing Sequences in the HP Model. *Journal of Biological Physics* 28:1-15 (2002).
95. Jeffery CJ. Moonlighting proteins. *Trends Biochem Sci* 24:8-11 (1999).

96. Jensen RA. Enzyme recruitment in the evolution of new functions. *Annu Rev Microbiol* 30:409-425 (1976).
97. Karchin R, Kelly L and Sali A. Improving functional annotation of non-synonomous SNPs with information theory. *Pac Symp Biocomput* 2005:397-408 (2005).
98. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS and Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18 (2002).
99. Kauzmann W. Some factors in the interpretation of denaturation. *Advances in Protein Chemistry* 14:1-57 (1958).
100. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H and Phillips DC. The three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181:662-666 (1958).
101. Khersonsky O and Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471-505 (2010).
102. Khersonsky O, Roodveldt C and Tawfik DS. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* 10:498-508 (2006).
103. Kimura M. Evolutionary rate at the molecular level. *Nature* 217:624-626 (1968).
104. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press (1983).
105. Koonin EV, Wolf YI and Karev GP. The structure of the protein universe and genome evolution. *Nature* 420:218-223 (2002).
106. Kriventseva EV, Biswas M and Apweiler R. Clustering and analysis of protein families. *Current Opinion in Structural Biology* 11:334-339 (2001).
107. Kunin V, Teichmann SA, Huynen MA and Ouzounis CA. The properties of protein family space depend on experimental design. *Bioinformatics* 21:2618-2622 (2005).
108. Kussell E. The designability hypothesis and protein evolution. *Protein and Peptide Letters* 12:111-116 (2005).
109. Ladunga I. Phylogenetic continuum indicates galaxies in the protein universe: preliminary results on the natural group structures of proteins. *J Mol Evol* 4:358-375 (1992).
110. Lau KF and Dill KA. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997 (1989).
111. Levinthal C. Are there pathways of protein folding? *J Med Phys* 65:44-45 (1968).
112. Levitt M. Nature of the protein universe. *Proc Nat Acad Sci USA* 106:11079-11084 (2009).
113. Li H, Tang C and Wingreen NS. Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix. *Proteins* 49:403-412 (2002).
114. Li H, Helling R, Tang C and Wingreen NS. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science* 273:666 (1996).
115. Li WH. *Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts (1997).

116. Lipman DJ and Wilbur WJ. Modelling neutral and selective evolution of protein folding. *Proc R Soc Lond B* 245:7-11 (1991).
117. Lomas DA and Carrell RW. Serpinopathies and the conformational dementias. *Nat Rev Genet* 3:759-768 (2002).
118. Lu Y and Freeland S. On the evolution of the standard amino-acid alphabet. *Genome Biology* 7:102 (2006).
119. Lupas AN, Ponting CP and Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion or relics of an ancient peptide world? *Journal of Structural Biology* 134:191-203 (2001).
120. Mann M. Computational methods for lattice protein models. PhD Thesis. Freiburg University. (2011).
121. Marti-Renom AM, Stuart AC, Fiser A, Sanchez R, Melo F and Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291-325 (2000).
122. Matias Rodrigues JF and Wager A. Evolutionary plasticity and innovations in comples metabolic networks. *PloS Comput Biol* 5(12):e1000613 (2009).
123. Maynard Smith J. Natural selection and the concept of a protein space. *Nature* 225:563-564 (1970).
124. McAuliffe JD, Pachter L and Jordan MI. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics* 20:1850-1860 (2004).
125. Mélin R, Li H, Wingreen N, and Tang C. Designability, Thermodynamic Stability, and Dynamics in Protein Folding: a Lattice Model Study. *J Chem Phys* 110, 1252 (1999).
126. Morgan II JC. Point set theory. Monographs and Textbooks in pure and applied mathematics (1989).
127. Murzin AG, Brenner SE, Hubbard T and Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-540 (1995).
128. Namba K. Roles of partly unfolded conformations in macromolecular selfassembly. *Genes to Cells* 6:1-12 (2001).
129. Needleman SB and Wunsch CD. A general method applicable to the search of similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443-453 (1970).
130. Nei M. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution* 22:2318-2342 (2005).
131. Nishikawa K. Island hypothesis. Protein distribution in the sequence space. *Viva Origino* 21:91-102 (1993).
132. Noirel J and Simonson T. Neutral evolution of Protein-protein interactions: a computational study using simple models. *BMC Structural Biology* 7:79 (2007).
133. Ohno S. *Evolution by gene duplication*. Springer-Verlag, Berlin (1970).
134. Orengo CA and Taylor WR. A rapid method of protein structure alignment. *J Mol Biol* 147:517-551 (1990).
135. Orengo CA and Thornton JM. Protein families and their evolution-a structural perspective. *Annu Rev Biochem* 74:867-900 (2005).

136. Orengo CA, Flores TP, Taylor WR and Thornton JM. Identification and classification of protein fold families. *Protein Eng* 6:485-500 (1993).
137. Orengo CA, Jones DT and Thornton JM. Protein superfamilies and domain superfolds. *Nature* 372:631-634 (1994).
138. Orengo CA, Silliote I, Reeves G and Pearl FMG. What can structural classification reveal about protein evolution? *Journal of Structural Biology* 134:145-165 (2001).
139. Patthy L. Genome evolution and the evolution of exon shuffling – a review. *Gene* 238:103-114 (1999).
140. Perelson AS and Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* 81:645-670 (1979).
141. Ptitsyn OB and Volkenstein MV. Protein structure and neutral theory of evolution. *J Biomol Struct Dyn* 4:137-156 (1986).
142. Qian J, Luscombe NM and Gerstein M. Protein families and fold occurrence in genomes: power-law behaviour and an evolutionary model. *J Mol Biol* 313:673-681 (2001).
143. Raes J, Harrington ED, Singh AH and Bork P. Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol* 17:362-369 (2007).
144. Raman K and Wagner A. The evolvability of programmable hardware. *Journal of the Royal Society Interface* 8:269-281 (2011).
145. Reeves GA, Dallman TJ, Redfern OC, Akpor A and Orengo CA. Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* 360:725-741 (2006).
146. Reidys C, Stadler PF and P Schuster. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull Math Biol*. 59:339-397 (1997).
147. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 12:85-94 (1999).
148. Rost B. Enzyme function less conserved than anticipated. *J Mol Biol* 18:595-608 (2002).
149. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M and Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32: 5539–5545 (2004).
150. Sander C and Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56-68 (1991).
151. Sali A and Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 234:779-815 (1993).
152. Sali A, Shakhnovich E, Karplus M. How does a protein fold. *Nature* 369: 248-251 (1994).
153. Salzberg SL, White O, Peterson J and Eisen JA. Microbial Genes in the Human Genome: Lateral Transfer or Gene Loss? *Science* 292:1903-1906 (2001).
154. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S et al. Database

resources of the national center for biotechnology information. *Nucleic Acids Research* 39:D38-D51 (2011).
155. Schuster P, Fontana W, Stadler P and Hofacker IL. From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc London B* 255:279-284 (1994).
156. Seitz CL. The cosmic cube. *Communications of the ACM* 28:22-33 (1985).
157. Sela M, White FH and Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. *Science* 125:691-692 (1957).
158. Silverstein KAT, Shoop E, Johnson JE and Retzel EF. MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics* 17:249-261 (2001).
159. Smith TF and Waterman MS. Comparison of biosequences. *Adv App Math* 2:482-489 (1981).
160. Sonnhammer ELL, Eddy SR, Birney E, Bateman A and Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* 26:320-322 (1998).
161. Sumedha, Martin OC and Wagner A. New structural variation in evolutionary searches of RNA neutral networks. *Biosystems* 90:475-485 (2007).
162. Szilagyi A, Gyorffy D and Zavodszky P. The Twilight Zone between Protein Order and Disorder. *Biophys J* 95:1612–1626 (2008).
163. Tatusov RL, Koonin EV and Lipman DJ. A genomic perspective on protein families. *Science* 278:631-637 (1997).
164. Taverna DM and Goldstein RA. Why are proteins marginally stable? *Proteins* 46:105-109 (2002).
165. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* 38:D142-D148 (2010).
166. Todd AE, Orengo CA and Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307: 1113–1143 (2001).
167. Tokuriki N and Tawfik D. Stability effects of mutations and protein stability. *Current Opinion in Structural Biology* 19:1-9 (2009a).
168. Tokuriki N, Stricher F, Serrano L and Tawfik D. How protein stability and new functions trade off. *PloS Comp Biol* 4:e1000002 (2008).
169. Waddington CH. Selection of the genetic basis for an acquired character. *Nature* 169:278 (1952).
170. Wagner A. Does evolutionary plasticity evolve? *Evolution* 50:1008:1023 (1996).
171. Wagner GP and Altenberg L. Complex adaptations and the evolution of evolvability. *Evolution* 50:967-976 (1996).
172. Wagner G and Wuthrich K. Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitior. *J Mol Biol* 130:31-37 (1979).
173. Wang X, Minasov G and Shoichert BK. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol* 320:85-95 (2002).

174. Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng* 11:621-626 (1998).
175. Waterman MS. *Introduction to computational biology*. Chapman & Hall, London (1995).
176. Whitehead DJ, Wilke CO, Vernazobres D and Bornberg-Bauer E. The look-ahead effect of phenotypic mutations. *Biology Direct* 3:18 (2008).
177. Williams PD, Pollock DD and Goldstein RA. Evolution of functionality in lattice proteins. *J Mol Graph Model* 19:150–156 (2001).
178. Williams PD, Pollock DD and Goldstein RA. Functionality and the evolution of marginal stability in proteins. *Evolutionary Bioinformatics* 2:1-11 (2006).
179. Wingreen NS, Li H and Tang C. Designability and thermal stability of protein structures. *Polymer* 45: 699-705 (2004).
180. Wolf YI, Grishin NV and Koonin EV. Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 299:897–905 (2000).
181. Wolynes PG. As simple as can be? *Nat Struct Biol* 4:871-874 (1997).
182. Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones D.F. (ed.), *Proceedings of the Sixth International Congress of Genetics, Brooklyn Botanic Garden*, Brooklyn, NY (1932).
183. Wroe R, Bornberg-Bauer E and Chan HS. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys J* 88:118–131 (2005).
184. Wroe R, Chan HS and Bornberg-Bauer E. A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J*. 1:79-87 (2007).
185. Xia Y and Levitt M. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA* 99:10382-10387 (2002).
186. Xia Y and Levitt M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* 55:107-114 (2004).
187. Xu YO, Hall RW, Goldstein RA and Pollock DD. Divergence, recombination and retention of functionality during protein evolution. *Human genomics* 2:1-10 (2005).
188. Yahyanejad M, Kardar M and Tang C. Structure space of model proteins: A principal component analysis. *J Chem Phys* 118:4277-4285 (2003).
189. Zhang CT. Relations of the numbers of protein sequences, families and folds. *Protein Eng* 10:757-761 (1997).
190. Zhang C and DeLisi C. Estimating the number of protein folds. *J Mol Biol* 284:1301–1305 (1998).
191. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E and Skolnick J. On the origin and highly likely completeness of single-domain protein structures. *Proc Natl Acad Sci USA* 103:2605-2610 (2006).

192. Zhou T, Drummond DA and Wilke CO. Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Evol* 66(4):395-404 (2008).
193. Zuckerkandl E and Pauling L. Evolutionary Divergence and Convergence in Proteins. Vernon Bryson and Henry Vogel, eds., *Evolving Genes and Proteins.* New York: Academic Press, pp. 97–166 (1965).
194. Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9: 133–148 (1981).

Acknowledgements

Firstly, I would like to thank those that without their help, this dissertation would not have been possible. I thank my Mother, Gladys Poblete, and my friends: Pablo Ramdohr, Ismael Vergara, Nicolas Zlatar, Margot Crucet, Noelia and Rosalino Rodriguez, Natalia Carrasco, Tomas Norambuena, Razvi Farooque, Anshumali Mittal and Rudolf Sagesser.

Secondly, I would like to thank those that made this dissertation what it is, and that without them, it would have been different. I thank Andreas Wagner for the opportunity to work in his lab. I thank former and current members of the Wagner Lab, especially those that I had the opportunity to collaborate scientifically with: Carlos Espinosa-Soto, Eric Hayden, Sorcha McGinty, Saurabh Pophaly, Daniel Rankin and Niv Sabath.

I also thank Martin Mann, the scientific committee of this dissertation and the external reviewer, Francisco Melo Ledermann. They provide insightful comments that help to improve this dissertation.

Finally, I would like to thank the MLS Program and the Forschungskredit of the University of Zurich for financial support.